

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
1 February 2001 (01.02.2001)

PCT

(10) International Publication Number
WO 01/08039 A2

- (51) International Patent Classification⁷: G06F 17/00
- (21) International Application Number: PCT/US00/20401
- (22) International Filing Date: 27 July 2000 (27.07.2000)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
09/361,122 27 July 1999 (27.07.1999) US
- (71) Applicant (for all designated States except US): ZENECA LIMITED [GB/GB]; 15 Stanhope Gate, London W1Y 6LN (GB).
- (72) Inventor; and
- (75) Inventor/Applicant (for US only): LERMAN, Charles, L. [US/US]; 501 Bishop Hollow Road, Newton Square, PA 19073-3138 (US).
- (74) Agents: BIRD, Donald, J. et al.; Pillsbury Madison & Sutro, LLP, 1100 New York Avenue, N.W., Washington, DC 20005 (US).
- (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.
- (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- Published:
— Without international search report and to be republished upon receipt of that report.
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/08039 A2

(54) Title: ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL DATA SETS USING LOW-RESOLUTION DATA GROUPING

(57) Abstract: Methods, systems and devices for operating on data provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and apply at least one of the grouping rules to the data. The data may be in a table, wherein the at least one grouping rule applies to at least one user-selectable column of the table. The grouping rule defines breakpoints corresponding to the user-definable number of groups, and application of the at least one rule to the data divides the data into groups based on the breakpoints. The grouped data is presented in a manner that visually distinguishes the groups, sometimes by coloring an aspect of the data according to the rules.

Best Available Copy

**ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL
DATA SETS USING LOW-RESOLUTION DATA GROUPING**

A portion of the disclosure of this patent document contains material
5 which is subject to copyright protection. The copyright owner has no objection to
the facsimile reproduction by anyone of the patent document or the patent
disclosure, as it appears in the Patent and Trademark Office patent file or records,
but otherwise reserves all copyright rights whatsoever.

10 **1. Field of the Invention**

This invention relates to analysis and pattern recognition of data. More
particularly, this invention relates to methods, systems and devices and
combinations thereof for analysis and pattern recognition in large sets of
multidimensional data using low-resolution data grouping.

15 **2. Background**

With the advent of computerization and the low cost of data storage and
acquisition, people in many endeavors are now accumulating very large sets of
data. For example, scientists in drug and chemical companies now use automation
to perform so-called high-throughput screening ("HTS") of chemical compounds.
20 HTS uses automated, relatively low-cost techniques to obtain various items of
information about chemical compounds. The goal of using HTS is to obtain
information about a very large number of compounds in a quick and relatively
low-cost manner. Having accumulated a very large HTS data set, it is necessary
to evaluate the data in order to determine which, if any, of the analyzed

compounds warrants further investigation. However, the results of such HTS tend to be very large sets of multidimensional data, on the order of thousands of rows and dozens of columns, and so it is very difficult to make decisions just by looking at the data.

5 In addition to the very large amounts of data produced by HTS, difficulties in existing data handling and analysis methods include the following:

- Data comes from very diverse sources, including HTS laboratories, physical measurements, bio-scientists' laboratories, various computational
10 software programs, etc., and the different sources tend to have very diverse kinds of output including numbers, text, mixed data types, error notations, blank data, replicate data (more than one value per compound), etc.
- Not all sources produce data on the same list of compounds, or in the same order.
- 15 • Some data values are misleadingly too precise, i.e., have high relative experimental errors or noise, and can easily be over-interpreted.
- Medicinal chemists have to weigh very different kinds of factors (for example, molecular weight vs. dose-responsiveness vs. ClogP vs. secondary biology vs. selectivity across screens) in trying to determine
20 which are the best compounds or clusters of compounds to which to devote further work.

SUMMARY OF THE INVENTION

This invention solves the above and other problems by providing automated tools to help with and speed up these data handling and analysis processes. These tools embody some assumptions about how the data should be
5 treated by internalizing the most generally acceptable assumptions, but leaving more idiosyncratic decisions to individual users.

A central concept on which this invention is based is grouping data into a relatively small number of categories using low-resolution data grouping. The grouping is visualized by assigning colors to data groups, e.g., in spreadsheets.
10 Grouping of data potentially changes the precision of the data.

This categorization of data has several major benefits, including:

- creating a visual means of finding data patterns;
- beneficially blurring small variations in numerical data that are, in practice, excessively fine distinctions, possibly due to experimental
15 error; and
- providing, in the colors themselves, a means or “common currency” to evaluate candidates across a wide range of data types.

Accordingly, in one aspect, this invention provides mechanisms to expedite pattern recognition in large sets of multidimensional data, such as those
20 that chemists assemble when evaluating hits from high-throughput screening (HTS) and deciding which ones will get priority for further investigation. In controlled trials, this invention has reduced the time to evaluate real data sets, from days of intense human effort, which is vulnerable to errors due to volume or fatigue, to a few minutes of automation with graphical presentation of results.

It quickly becomes obvious upon using the system that the tools also have value in data-handling areas other than HTS. Examples include selection and management of any kind of tabulated data, e.g., portfolio management for any kind of rated portfolios, selection of drug candidate compounds, selection and management of proteins that are candidates for targets for drugs, selection and management of research projects competing for resources, and evaluating employee performance or job candidates.

The system of this invention includes a new special command menu, a set of graphical user interface worksheets, and action buttons to facilitate the coloring and color analysis processes for the user. While the central process is the data grouping and coloring, there are also new tools for the upstream, or pre-grouping and coloring processes of importing, assembling, regularizing, and characterizing data in a spreadsheet, and for the downstream processes of visualizing, scoring, comparing, and sorting large amounts of color-coded data. The data-grouping and spreadsheet-coloring tool is presently implemented with a flexible, powerful, and convenient user interface that does not require knowledge of spreadsheet macros or of the Visual Basic language (used for the system's implementation).

Accordingly, this invention provides methods, systems and devices for operating on data.

In one aspect, the method of this invention provides at least one user-defined grouping rule for grouping the data into a user-definable number of groups. At least one of the grouping rules is applied to the data. The data may be provided in a table and the grouping rule applies to at least one user-selectable column of the table. In some embodiments, the grouping rule defines breakpoints corresponding to the user-definable number of groups. Application of the rule the

data divides the data into groups based on the breakpoints. The method may include presenting the grouped data in a manner that visually distinguishes the groups. In some embodiments, the grouping rules associate colors with groups and the grouped data is presented with an aspect of the data colored according to the rules.

Sometimes the data are in labeled columns in a spreadsheet, and the grouping rule specifies at least one breakpoint and a corresponding color for each range defined by the breakpoint. The grouped data are presented by coloring each data item in one labeled column of the data based on the breakpoint and the corresponding color of the breakpoint.

The breakpoints may be numeric or textual values. In some embodiments, the breakpoint is determined automatically based on the data.

Sometimes the data are provided in a table, and backgrounds of table cells are colored according to the rules.

The number of groups may be fewer than a number of possible data values.

In another aspect, this invention is a method of operating on data by providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups. The grouping rule is applied to the data to generate grouped data. At least one user-defined scoring rule is used to score grouped data according to user-defined scores. The scoring rule is applied to the grouped data to score the grouped data.

In yet another aspect, this invention is a method of operating on data, in which data are grouped by applying to the data at least one user-defined grouping rule for grouping the data into a user-definable number of groups. The grouped

data are scored by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

In some embodiments the data can be a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be
5 used in scoring the cases. The scoring applies the function to the data to obtain a score for each case. Sometimes the method includes sorting the scored cases by score, individually or by cluster, as described below.

The notion of clustering is that subsets of the various cases may be
10 associated into clusters by having identical entries in any user-selected column of data, known as a clustering column. In some embodiments of the invention, the integrated clusters are treated by averaging the properties of all the cases which comprise each cluster.

Thus, according to aspects of this invention, in order to facilitate analysis
15 and pattern recognition in large, multidimensional data sets, the precision of the data is potentially changed (implemented, e.g., by grouping the data) and then the data are presented for visualization (implemented, e.g., by coloring the data).

BRIEF DESCRIPTION OF THE DRAWINGS

This file contains at least one drawing executed in color. Copies of this
20 patent with color drawings will be provided by the United States Patent and Trademark Office upon request and payment of the necessary fee.

The above and other objects and advantages of the invention will be apparent upon consideration of the following detailed description, taken in

conjunction with the accompanying drawings, in which the reference characters refer to like parts throughout and in which:

FIGURE 1 shows a typical computer system on which the present invention operates;

5 **FIGURE 2** shows an overview of the functionality of the present invention;

FIGURES 3A-3B depict a display of data in a spreadsheet;

FIGURES 4A-4B show a color control rules worksheet according to one embodiment of the present invention;

FIGURES 5A-5B show data coloring rules;

10 **FIGURES 6A-6C** show a data coloring control panel and a flow chart of the data coloring process, respectively;

FIGURES 7A-8B show the worksheet of **FIGURE 3A** and **3B** after various coloring rules in **FIGURE 4A** have been applied;

FIGURES 9A, 9B, 10A, and 10B depict displays of data in spreadsheets;

15 **FIGURES 11A and 11B** show the form of the cluster control worksheet according to one embodiment of the present invention;

FIGURES 11C-11D shows control panels from the cluster control worksheet of **FIGURES 11A-11B**;

20 **FIGURE 12** shows the enlarging of the cluster starts mechanism according to one embodiment of the present invention;

FIGURES 13A-13D show the application of vertical display re-scaling according to one embodiment of the present invention;

FIGURES 14A-14D and 15A-15B show the application of the scoring and sorting of clusters according to one embodiment of the present invention;

FIGURES 16A-16N, 16P and 16Q show aspects of the application of the dose-response scoring and estimation of potencies according to one embodiment of the present invention;

FIGURES 17A-17B show the application of the sheet statistics tool
5 according to one embodiment of the present invention;

FIGURES 18A-18D show the application of the scoring and sorting of clusters for the purpose of project prioritization and management according to one embodiment of the present invention;

FIGURES 19-24 show examples of the application of this invention to
10 various types of data; and

FIGURES 25 and 26 show application of an aspect of this invention.

DETAILED DESCRIPTION OF THE PRESENTLY PREFERRED EXEMPLARY EMBODIMENTS

Overview

FIGURE 1 shows a typical computer system **100** on which the present
15 invention operates. The computer system **100** includes a processor (CPU) **102** connected to a memory system **104** and a display **106**. The computer system also includes various input devices including a keyboard **108** and a mouse **110** or other pointing device. Internal storage **112** (e.g., a hard disk, a CD ROM and the like)
20 and external storage **114** (such as a floppy disk, CD ROM and the like) are also provided.

Various aspects of this invention are implemented as computer software programs or algorithms **116** which run on the computer system **100**. The software programs **116** can reside in the internal storage **112**, the external storage **114**,

and/or in the memory 104. The software programs 116 operate on data 118 which is provided, e.g., on the external storage 114. The software programs 116 operate in a standard and known manner by being executed on the processor 102 of the computer system 100.

5 In some embodiments of the present invention, the user can create and modify various executable rules 120 which can operate on the data 118. For the sake only of explanation, the rules 120 are depicted separately from the data in the figures. However, as explained in more detail below, some or all of the rules 120 can be part of the data 118.

10 In preferred embodiments, the computer system 100 is capable of running the spreadsheet program EXCEL™ 95 (hereinafter "EXCEL") from Microsoft Corporation, and the software computer programs 116 are written in Microsoft Corporation's Visual Basic (hereinafter "VB") and are provided as an add-in to EXCEL. A single copy of software thus serves all data files on a particular
15 machine. To conserve EXCEL resources, in some embodiments, the package self-installs the add-in when the user opens a data file, and un-installs the add-in when the last data file in memory is closed.

 In a preferred embodiment, this invention works entirely within the environment of EXCEL. EXCEL structures data files as workbook files which
20 contain worksheets. The programs 116 of this invention consist of special EXCEL worksheets, called control sheets, on which input data is written by the user into designated labeled cells. The control sheets are part of the same EXCEL workbook file as the data. The control sheets also contain action buttons to execute the various procedures associated with this invention. The rules 120 are formed by
25 setting various parameters in the control sheets.

When the workbook file is saved, the parameters (for the rules 120) are stored on the control sheets along with the data, and they can be modified and/or re-executed at any time without having to re-enter anything. The results of operations are automatically written as worksheets in the same workbook file,
5 providing a convenient, integrated data environment in a single file.

The system according to the present invention operates, in one aspect, in accordance with **FIGURE 2**. Recall that the user's aim is to perform analysis and pattern recognition in large, multidimensional data sets using (potentially low resolution) data grouping. To this end, the user and/or the system will create rules
10 for coloring and presenting the data. First (at 122) a user creates and organizes the data 118. Various tools (discussed below) are provided to aid in the creation and organization of the data. Then (at 124) the user creates rules 120 for operating on the data 118. The rules 120 can be created before or after the data 118, rules can be reused for different sets of data and multiple rules can apply to the same data.
15 The creation and operation of rules are discussed in greater detail below. Once the data 118 and the rules 120 are created, the user then selects some (or all) of the rules to apply to the data (at 126). Specifically, the user groups and thereby colors the data according to selected rules. With the data grouped and colored according to the rules, the user can then perform group/color-mediated data mining (at 128).

20 **FIGURES 3A-3B** show views of the program of this invention in operation with a sample EXCEL sheet 300, denoted "DEMO 1" (302) containing data (not all the data in the sheet is visible). The views of EXCEL worksheets shown in the various figures and examples that follow are the views that are presented on the display 106 of the computer system 100. Sheets in an EXCEL workbook are
25 labeled with tabs at the bottom of the worksheet. The data on the "DEMO1" sheet

300 consists of eight columns of data for each of a number of compounds. The compounds are denoted "Cmpdxx", where "xx" ranges from "01" to the number of compounds. In FIGURE 3B, the last compound visible on the data sheet is "Cmpd58". The eight columns are headed:

- 5 1. "Cmpd" (column A);
2. "Series" (column B);
3. "Test1" (column C);
4. "Test2" (column D);
5. "Test3" (column E);
- 10 6. "HTS SPA Dose-Resp % Inhib @3x10-6M" (column F);
7. "HTS SPA Dose-Resp % Inhib @ 1x10-6M" (column G);
8. "HTS SPA Dose-Resp % Inhib @ 3x10-7M" (column H); and
9. "HTS SPA Dose-Resp % Inhib @ 1x10-7M" (column I).

In addition to the "DEMO 1" worksheet 300, the EXCEL workbook shown
15 in FIGURES 3A and 3B has seven other worksheets, denoted "DEMO 2" 304;
"DEMO 3" 306; "clusterinfo DEMO" 308; "Append Control" 310; "Color
Control" 312 and "Cluster Control" 314. The last three worksheets, denoted
respectively "Append Control"; "Color Control" and "Cluster Control," contain
various rules and controls (to be discussed below). The data in worksheets
20 denoted "DEMO 1" 302; "DEMO 2" 304; "DEMO 3" 306; and "clusterinfo
DEMO" 308 correspond to data 118 (FIGURE 1) and the controls or rules in the
worksheets denoted "Append Control" 310; "Color Control" 312, and "Cluster
Control" correspond to the rules 120 (FIGURE 1).

FIGURES 4A-4B show a color control rules worksheet (312, denoted "color
25 control") according to the present invention, as displayed on display 106 of the

computer system 100. The color control worksheet 312 is shown with some rules already in place, i.e., having values set, and other rules left blank. These rules are shown as examples only, and, as with any of the other types of rules, any or all of the rules can be set by the user. A typical data coloring rule 130 is shown in

5 **FIGURE 5A.** The rule 130 has already been set up and operates on the appropriate data when selected by a user (using mouse 110, **FIGURE 1** or some other pointing device) in the area 132 marked "Click here to run these". The rule 130 (as with all of the preferred color control rules) has four parts, namely the name of the sheet 134 containing the data on which the rule is to operate ("DEMO 1" in the example

10 of **FIGURE 5A**); the columns 136 of data of the sheet on which the rule is to operate ("E" in the example of **FIGURE 5A**); the number of colors 138 to be used by the rule 130; a number of breakpoints 140 (denoted "break 1" to "break 4" in the example of **FIGURE 5A**); and a corresponding number of colors 142 for each range defined by the breakpoints (denoted "color 1" to color 4" in the example of

15 **FIGURE 5A**). Actually, as explained below, the number of breakpoints is one less than the number of colors.. In the specific example shown in **FIGURE 5A**, the rule has three breakpoints of 1, 5 and 10, defining four ranges with four corresponding colors 142, namely light green, yellow, orange and red. Preferably the named colors are also depicted in the actual colors, so that, in this example, the

20 background of the word "lightgreen" is shown in light green, the background of the word "yellow" is shown in yellow and so on.

In this invention it is preferable to show data and meta-data (headings etc.) in color. In some embodiments, the coloring is implemented by showing a background area of the text representing the data in the appropriate color.

25 Sometimes the actual text representing the data is shown in the appropriate color.

In presently preferred embodiments, the font color is only changed in cases where necessary to improve contrast with the background color for readability. Only two font colors, dark (black) and light (pale gray), are used in the presently preferred embodiment. Combinations of both approaches can be used. For example, the background section of the word "yellow" is preferably shown in the color yellow. It is also possible to show the word itself, i.e., the font, in the color yellow, as long as that color is distinguishable from the background.

The particular rule 130 shown in FIGURES 4A and 5A, operates as follows, when selected:

In sheet "DEMO 1" 302, in column E, values less than or equal to 1 (break 1) are colored light green (color 1); values in the range 1 to 5 (between break 1 and break 2) are colored yellow (color 2); values in the range 5 to 10 (break 2 to break 3) are colored orange (color 3); and values greater than 10 (break 3) are colored red (color 4).

Another typical data coloring rule 130-1 from the color control sheet 312 is shown in FIGURE 5B. The rule 130-1 is set up to operate on columns "C" and "D" of sheet "DEMO 1". The rule 130-1 uses three (3) breakpoints (break1=0.1, break2=1 and break3=5) defining four ranges with four (4) corresponding colors ("lightgreen", "yellow", "orange", and "red").

The results of applying the rule 130-1 of FIGURE 5B to the data in sheet "DEMO 1" (302, FIGURE 3) are shown in FIGURES 7A-7B. As can be seen from FIGURES 7A-7B, after application of the rule 130-1, all of the data in columns C and D of the sheet labeled "DEMO 1" has been colored according to the rule. Specifically, data having a value less than or equal to break 1 (0.1) have been colored light green; data values in the range between break 1 and break 2 (0.1 to

1) have been colored yellow; data values in the range between break 2 and break 3 (1 to 5) have been colored orange; and data values greater than break 3 (5) have been colored red.

The results of applying all of the other color control rules shown in
5 **FIGURES 4A-4B** to the data in sheet "DEMO 1" are shown in **FIGURES 8A-8B**.
The rules can be applied individually (as shown above with respect to
FIGURES 7A-7B), or they can be all be applied at the same time. In order to apply
all rules to a particular data set (sheet), each rule can be individually selected or
the area labeled "RE-RUN ALL RULES FOR SHEET NAMED DEMO 1" (on the right
10 side of **FIGURE 4A**) can be selected. Note that if two rules apply to the same
column of the same sheet, the second rule run on that column will override the
first rule run on that column.

To create a coloring rule a user performs the following (with reference to
FIGURE 6B):

- 15 (1) Select the "COLOR CONTROL" sheet 312 and pick a control panel on
that sheet to use (an empty panel or one containing a rule no longer
needed) (at 600). All control panels on a sheet can be cleared by
clicking the button labeled "CLEAR ALL ENTRIES ON THIS SHEET"
(318 in **FIGURE 4A**).
- 20 (2) In the selected control panel, enter the name of the sheet to be
colored (at 602).
- (3) In the selected control panel, enter a column or columns (at 604).
For multiple columns, either list them separated by commas, or use
a colon or hyphen to denote ranges, or some combination. For
25 example, "A:D,F" means columns A,B,C,D, and F. To aid in

choosing columns, the user can right-click on the cell containing the name of the data sheet, and pick "Open Twin Screen" from the shortcut menu that appears, to create a special dual display. This also creates a "Close Twin Screen" button to go back.

- 5 (4) Choose a number of colors to use (at 606), either by entering the number of colors or by repeatedly clicking the gray button adjacent the cell labeled "# of colors". In preferred embodiments, the system allows for five breakpoints and six colors per rule. Accordingly, the numbers will cycle from 1 to 6, and various cells
10 below them will be blacked out accordingly.

- (5) Enter the breakpoints that define the color groups (at 608), in any of three modes:

- a) Numeric data, manual mode: enter numbers to form the breakpoints, i.e., the boundaries between the
15 color groups, one less than the number of colors, in increasing numerical order. Cells whose values exactly equal a breakpoint value will be colored with the lower group (breakpoint 1 is colored with color 1, etc.)

- 20 b) Numeric data, automatic mode: enter either "value", "log", or "count" as the first breakpoint. If multiple columns have been chosen, the user must also enter "yes" or "no" opposite "Re-scale all?" at the bottom of the panel, to indicate whether each column
25 should get its own auto-breakpoints, or whether the

auto-breakpoints of the first column (first in list in the rule, not first on the data sheet) should be used for all.

This mode reports information about the breakpoints it determines, and thus could also be used to explore the distribution of numerical values in a column prior to a final *manual* breakpoint selection.

- c) Text data: enter the strings to be matched and colored, in preferred embodiments, up to five (5) in number. Matching is case-insensitive unless the string is enclosed in double quotes (“ and ”); otherwise, no quotation marks are necessary.

Several special text strings act as operators if entered as the first word in a rule cell:

<u>Rule Entry</u> (OPERATORS need not be uppercase—here only for emphasis)	<u>Meaning</u>
test string	color data cell if its whole content matches the test string
NOT test string	<ul style="list-style-type: none"> • color data cell if its whole content does not match the test string; • will not color numeric cells
CONTAINS test string	color data cell if contains the test string as a substring anywhere
NOTCONTAINS test string	<ul style="list-style-type: none"> • color data cell if it does not contain the test string anywhere; • will not color numeric cells

<u>Rule Entry</u> (OPERATORS need not be uppercase—here only for emphasis)	<u>Meaning</u>
BEGINS test string	color data cell if it begins with the test string
ENDS test string	color data cell if it ends with the test string
* (an asterisk)	(wildcard) color data cell containing any data, including numeric cells
BLANK	color data cells that are blank

Using quotes to force matching to be case-sensitive also works with strings that follow an operator.

5 It is possible to construct a text-coloring rule in which certain cells may satisfy more than one of the “breakpoint” values. For example, if a rule says that “active” is colored green and “contains act” is colored red, then the word “active” in a cell would satisfy both. In such cases, the system colors the cell according to the first condition satisfied on the list of breakpoints. This dependence on the order can be used advantageously to achieve complex coloring conditions. The sequence of conditions can be considered as a series of filters, through which only the as-yet-uncolored cells fall through to the next decision.

10

15

- (6) Enter the names of the colors to use (at 610), in the order corresponding to the breakpoints. A display of color samples is provided at the right side of the Color Control sheet 312. A user need only enter the name, and the appropriate cell will become colored when the tool is executed. If the user wants the color to display immediately, he can copy and paste the sample cell into the rule's color cell. A special pseudo-color named "SKIP" is used to tell the system not to color the cells whose data falls in this group.
- (7) When the rule has been created, the user executes the rule by selecting the rule's "CLICK HERE TO RUN THESE" button on the panel filled in (at 612, FIGURE 6C).
- (8) To create different coloring rules for other columns, repeat the above in additional control panels. If the user runs out of control panels, he can create more control panels by copying an existing one and pasting it onto a blank section of the color control sheet.

To the extent that a single panel cannot hold all the requirements for a particular rule, a user can combine two or more panels to create a particular rule. For example, if a user needs ten (10) breakpoints, two panels can be used.

- With reference to the coloring rule is shown in FIGURE 6A, once the rule has been set, a number of parameters are stored in the system. The parameters are "sheet name" ("DEMO 3" in FIGURE 6A), column specification, number of colors, array of breakpoints, array of colors, and multicolumn scaling mode.

The data coloring mechanism operates as follows, with reference to the flowchart of **FIGURE 6C.**:

1. The user enters the parameters into a rule panel on a "COLOR CONTROL" worksheet **312**, e.g., as described above with reference to the panel of **FIGURE 6A.**
2. The user selects (clicks) the activation button (labeled "Click here to run these") on that panel (at **612**). This causes the system to:
 - (A) Read and interpret the parameters from the panel (at **614**).
The system can identify which button was clicked using the Visual Basic ("VB") "caller" property, described in more detail below. The parameters are then read based on the identity of the cell location of the button using the VB "TopLeftCell" property. The system retrieves the parameters (sheet name, column specification, number of colors, array of breakpoints, array of colors, and multicolumn scaling mode) from cells in this panel by relative reference to the button cell.
 - (B) Next, determine the mode of the coloring rule (at **616**) (i.e., numeric v. text or manual v. automatic, and, if automatic, which of value, log or count). This uses the analysis of the first breakpoint entry.
 - (C) Compile a list of the columns specified in the "column specification" parameter (at **618**). This is done by scanning the various areas contained in the selection, as follows:

```

For Each singlearea In Selection.Areas
  For Each c In singlearea.Columns
    If Not CountEmpty Then
      lr = LastRowInColumn(c.Column)
    End If
  
```



```

        If Not CountEmpty And lr = 0 Then
            'skip this empty column
            ncols = ncols - 1
        Else
            ' add this column to the list
5         icol = icol + 1
            colnumarray(icol) = c.Column
        End If
    Next c
Next singlearea

```

10

(D) If an auto-breakpoint mode is being used (determined at 620), analyze the data values to determine the breakpoints (at 622). This is done by:

- (i) Collecting statistics on the data distribution in each specified column; and
- (ii) Calculating the automatic break points for the appropriate mode. For example, the auto-value breakpoints are determined as follows:

15

```

20 If breakmode = "VALUE" Then
    interval = (maxvalue - minvalue) / ncolors
    break(0) = minvalue
    For ibreak = 1 To ncolors - 1
        break(ibreak) = break(ibreak - 1) + interval
25 Next ibreak

```

(iii) Displaying the results for user approval or cancellation.

(C) Loop through the cells in the chosen columns on the chosen worksheet (at 624).

30

(D) Compare each cell's value to the list of breakpoints (at 626). If the coloring rule is in text mode, use the meanings of the special breakpoint operators ("contains", "blank", asterisk "*"; or quotation marks).

5

(E) When a match is found, apply the appropriate color (at 628).

The code below illustrates the processes (D) and (E) for numeric breakpoints:

10

```

For Each cell In Range(Cells(StartColoringRow, colnum),
                        Cells(FinishColoringRow, colnum))
    cvalue = cell.Value
    colored = False
15    If IsNumeric(cell.Value) Then
        If Not IsEmpty(cell) Then
            ' (have to test both because EMPTY is numeric)
            For ibreak = 1 To ncolors - 1
                If cvalue <= break(ibreak) Then
20                    If Color(ibreak) <> "SKIP"
                        cell.Interior.ColorIndex =
                            Color(ibreak)
                        Call TextContrast(cell)
                    End If
                colored = True
25                Exit For
            End If
            Next ibreak
        If colored = False Then
30            ' Not hit yet? Must be top category, so:
            If Color(ncolors) <> "SKIP"
                cell.Interior.ColorIndex = Color(ncolors)
                Call TextContrast(cell)

```

```
        End If
        colored = True
    End If
    End If
5      Else      ' not numeric - just don't color it
    End If
Next cell
```

The operation of the data coloring tool of this invention will now be

10 described in greater detail. Each coloring rule is provided in a coloring control panel that has the general form of a coloring rule as shown in **FIGURE 6A**. In one preferred embodiment, each coloring control panel **144** is a double-outlined unit, sixteen (16) cells high by two (2) cells wide. As noted above, a user is provided with coloring control panels on the color control worksheet **312**. A user can use

15 the coloring control panel **144** to set the sheet and column(s) on which the rule is to operate, the number of colors, the various break points and the colors associated with those breakpoints. The sheet is set by entering its name into the cell **146** adjacent the cell labeled "sheet". The column (or columns) on which the rule is to operate is (are) set by entering its (their) name in the cell **148** adjacent the cell

20 labeled "column(s)". The number of colors is set by the user by selecting the cell **150** adjacent to the cell labeled "# of colors". Each time the cell **150** is selected it increases the number of colors, up to a maximum of six (6), i.e., rotating through the values 1 to 6. I.e., when the cell **150** shows a "6" and is selected, it reverts back to "1". That is, selecting the cell **150** causes the value in the cell to cycle

25 from "1" to "6" and then back to "1".

A Visual Basic ("VB") macro function ("*CallColorColumn*") is associated with the top cell **152** of the control panel **144**. When the cell **152** is selected by

the user (with the mouse 110 or the like), the function associated with that cell is executed by the computer (CPU 102). In the presently preferred embodiments, the *CallColorColumn* function extracts the button name of the cell 152 and then calls a second function ("*CallColorColumnSubroutine*") with that button name as one of the parameters. The function *CallColorColumnSubroutine* takes three parameters, namely *ButtonName*, *StartColoringRow*, and *FinishColoringRow*. The two parameters *StartColoringRow*, and *FinishColoringRow* are optional.

First, the function *CallColorColumnSubroutine* determines what specific values to use for the coloring by reading them from the control panel 144. Since the values are all in fixed positions relative to the selected button cell 152 that initiated the call to the function *CallColorColumn*, the values can be determined once the location of that button cell 152 has been determined. This is done using the following Visual Basic code:

```

15      Sheets("Color Control").Activate
      headingrow =
          ActiveSheet.Buttons(ButtonName).TopLeftCell.Row
      headingcol =
          ActiveSheet.Buttons(ButtonName).TopLeftCell.Column
20

```

Note that if the function *CallColorColumnSubroutine* was called from another sheet (not "Color Control") then this method will not find it.

The various parameter values are then read as follows:

```

      Sheet name:
25      datacol = headingcol + 1
      sheetname = Trim(Cells(headingrow + 1, datacol).Value)

```

If there is no sheet named "sheetname" an error function is called.

Generally, in preferred embodiments, a great deal of error checking takes place to ensure that the user is provided with a friendly and useable interface to the program. Most of the error checking is not mentioned in this description,
5 however, one skilled in the art would know what kinds of error checking to implement in order to provide a user-friendly working environment.

The column(s) to be colored are specified by:

```
Cells(headingrow + 2, datacol).Value
```

The number of colors is specified by the variable *ncolors*, where:

```
ncolors = Cells(headingrow + 3, datacol).Value
```

15 Within the function *CallColorColumnSubroutine* there are two arrays, named *break* and *color*, which are used to store the breakpoints and colors, respectively. The first breakpoint is set as follows:

```
break(1) = Cells(headingrow + 4, datacol).Value
```

20 The value of the first breakpoint is used to determine the break mode ("NUMERIC", "VALUE", "LOG", or "COUNT"). If *break(1)* (as determined above) is numeric, then the mode is set to "NUMERIC", otherwise, if *break(1)* is one of "VALUE", "LOG", or "COUNT", then the break mode is set to that mode,
25 otherwise the break mode is set to "TEXT".

Next, the function determines whether multiple columns were specified, in which case it determines whether the user selected to re-scale all the columns.

The user's re-scale selection is determined by:

```

5      rescale_all_string = Cells(headingrow +
      15, datacol).Value

```

Now the rest of the breakpoints (if any) are read. If the break-mode is "AUTO" then the breakpoints are set as follows:

```

For i = 2 To lastbreaknum
10      break(i) = Cells(headingrow + 3 + i, datacol).Value

```

Various possible errors are checked for. E.g., if any breakpoints are missing (i.e., if *break(i)* is empty, the user is notified. Also, if the break mode is "NUMERIC" and non-numeric breakpoints are set, the user is notified. If
15 numeric breakpoints are not in increasing order, the user is notified. As noted above, generally, in preferred embodiments of the present invention, a great deal of error checking is performed on all user inputs to ensure that the values are correct and consistent. Most error checking is omitted from this description.

The *CallColorColumnSubroutine* maintains an array, *colorname*, which
20 maps integers to colors. In preferred embodiments, there are fifty six (56) colors available. To use the higher numbered colors, the computer's video card must be set appropriately. Using the *colorname* array, the program next associates the user provided color names with index numbers. Specifically, for each of the (up to six in a preferred embodiment) colors specified, the user specifies an actual color
25 name, denoted *cname*. This name is determined for each color by:

```

For j = 1 To ncolors
      cname = Cells(headingrow + 8 + j, datacol).Value

```

The interior of each color-specifying cell is then colored by the appropriate (selected) color by setting the color property (*Interior.ColorIndex*) of the cell:

```

Cells(headingrow + 8 + j, datacol).Interior.ColorIndex =
5      Color(j),

```

where the value of the variable *j* ranges from 1 to *ncolors*.

Then the cell is further processed by a function *TextContrast*.

```

      Call TextContrast(Cells(headingrow + 8 + j, datacol))
10      With the parameters read and checked, the system is ready to process and
color the selected sheet (specified at cell 146 in FIGURE 6A). The selected
columns (specified in cell 148 in FIGURE 6A) in the selected sheet are processed
one-by-one by the following program code:

```

```

Call ParseInput(InString, inspecifier)
15 For Each singlearea In Range(inspecifier).Areas
      For Each c In singlearea.Columns
          colnum = c.Column
          Call ProcessOneColumn(colnum, ncolors, break,
20                               Color, breakmode,
                               rescale_all, sheetname,
                               StartColoringRow,
                               FinishColoringRow)
      Next c
Next singlearea
25

```

The processing performed by the function *ProcessOneColumn* is as follows: The column designated by *colnum* on sheet *sheetname* is to be colored according to the breakpoints in the array *break* and the colors in the array *colors*. The designated column is colored from the row corresponding to

```

30 StartColoringRow to the row corresponding to FinishColoringRow. Note that the

```

function *ProcessOneColumn* is also provided with the break mode and the variable *rescale_all*.

Function *ProcessOneColumn* first calculates the automatic breakpoints, if necessary. Note that automatic breakpoints are determined from the whole
 5 column, even if this call says to color only a limited range of rows. If the value of *breakmode* is "VALUE" or "LOG" and the value of *rescale_all* is set to "True" Or the value of the first breakpoint (*break(1)*) is set to "VALUE" or "LOG", the program calls the function *AutoValueBreakpoints* as follows:

```
10      Call AutoValueBreakpoints(colnum, colletter, ncolors,
      break, Color, breakmode, rescale_all).
```

Otherwise, if the *breakmode* is set to "COUNT" and the value of *rescale_all* is set to "True" or the first breakpoint (*break(1)*) is set to "COUNT", then the program calls the function *AutoCountBreakpoints*, as follows:

```
15      Call AutoCountBreakpoints (colnum, colletter, ncolors,
      break, Color, breakmode, rescale_all, sheetname).
```

With the breakpoints calculated, the columns are colored according to the type of breakpoints specified by the user. Specifically, when the *breakmode* is any
 20 one of "VALUE", "COUNT", "LOG", or "NUMERIC", the system executes a function *ColorNumericColumn*. On the other hand, when the *breakmode* is "TEXT", the system executes a function *ColorNumericColumn*. The VB code for this is as follows:

```
25      Select Case breakmode
      Case "VALUE", "COUNT", "LOG", "NUMERIC"
```



```
        Call ColorNumericColumn(colletter, ncolors,  
break, Color, StartColoringRow, FinishColoringRow)  
        Case "TEXT"  
            Call ColorTextColumn(colletter, ncolors, break,  
5 Color, StartColoringRow, FinishColoringRow)  
        End Select
```

Then, when the rule in control panel **144** is selected for execution, the rule is applied to the selected column(s) (denoted in cell **148**) of the named sheet (in
10 cell **146**). For each column in the named sheet, the value in each cell is compared to the various breakpoints and the cell is colored corresponding to the appropriate breakpoint.

Examples of the application of various coloring rules in the "COLOR CONTROL" worksheet of **FIGURES 4A-4B**, are shown with reference to the data in
15 worksheet "DEMO 2" (depicted in **FIGURES 9A, 9B, 10A and 10B**).

Color-Mediated Data Mining

As noted above with reference to **FIGURE 2**, once the data have been colored according to the user-selected rules (at **126**), the user can then perform color-mediated data mining (at **128**). The presently preferred embodiment of this
20 invention provides five mechanisms (each discussed below) for color-mediated data mining, namely mechanisms to:

1. enlarge/shrink cluster starts;
2. vertically re-scale the display;
3. score and sort clusters; and
- 25 4. score and sort dose-response data.

The following discussion refers to the cluster control worksheet which is shown in **FIGURES 11A-11B**.

1. Enlarge/Shrink Cluster Starts

5 The “Enlarge Cluster Starts” mechanism highlights the first row of each cluster in clustered data by enlarging the font of the cell containing the cluster number or label, thus enabling size reduction of the spreadsheet for the user to focus on the color patterns. When the cell height is dramatically reduced in order to see more cells on a screen or printed page, this enlargement allows the user to
10 still read the label at the beginning of each cluster. The mechanism takes user input from a *Cluster Control* worksheet. A corresponding mechanism (“SHRINK CLUSTER STARTS”) allows for undoing the enlarging. This mechanism handles cluster numbers or textual labels. Any column can be designated as the cluster labels to be processed.

15 Operation of the mechanism is as follows:

- (1) From the “CLUSTER CONTROL” sheet **314** pick a control panel to use (one which is empty or one containing inputs no longer needed). On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input
20 information for several tools.
- (2) In the blue section, enter a sheet name and the column to be considered as the cluster labels.
- (3) Click either the blue-text “Enlarge Cluster Starts” or “Shrink Cluster Starts” button.

The program code accomplishes this by scanning the column of cluster labels, identifying any entries that are different from the one immediately above, and enlarging them. Code that carries out this function is shown below:

```

5  For Each c In Range(Cells(3, colnum), Cells(lastrow,
    colnum))
        irow = c.Row - 1
        icol = c.Column
        If c.Value <> Cells(irow, icol).Value Then
10         c.Font.Size = bigfontsize
            '      Rows(Irow + 1).RowHeight = bigrowheight
        End If
    Next c

```

Example

15 An example of the application of the enlarge cluster mechanism is shown in **FIGURE 12** which shows the application of a rule (shown in the control panel **FIGURE 11C**) from the cluster control worksheet in **FIGURE 11B** to the data of worksheet "DEMO 2" as shown after coloring in **FIGURES 10A-10B**. As shown in **FIGURE 11C**, the rule is to be applied to column B of sheet "DEMO 2".

20

2. Vertical Display Re-Scaling

The vertical re-scaling mechanism operates by taking a user-provided scale factor and then changing height of data rows to facilitate visualization of large-scale color patterns. The mechanism leaves column heading heights and column

25 widths unchanged. This makes headings remain readable and greatly simplifies examining long columns of data for color patterns.

FIGURES 13A-13D show the application of the vertical display re-scale mechanism according to the present invention. FIGURES 13A-13B show some of the data in the worksheet labeled "DEMO 3" 306 (FIGURE 13A shows the first thirty eight or so elements and FIGURE 13B shows the remaining elements of that worksheet). As can be seen from the figures, the worksheet "DEMO 3" 306 has three hundred and twenty eight (328) data entries (in rows 2-329). The user can vertically scale the display by selecting "Re-scale Vertical" from the system's special menu or by pressing a particular control key sequence (e.g., "CNTL-SHIFT-V" in a preferred embodiment). This presents the user with a dialog box 318, as shown in FIGURE 13C, which asks the user to enter a scaling factor relative to the current size. The user enters a scaling factor to enlarge or reduce or restore the display. In the example shown, the user enters a scaling factor of 0.1 which produces the vertically scaled display shown in FIGURE 13D.

Vertical scaling allows a user to get an overview of the data, based on the coloring.

The portion of the program code presented below carries out the central function of the vertical display rescaling mechanism:

```

rowspec = "2:" & lastrow ' leaves the headings unchanged,
i.e., readable
If factor = -1 Then
    Rows(rowspec).Rows.AutoFit
Else
    For irow = 2 To lastrow
        Rows(irow).RowHeight =
            Rows(irow).RowHeight * factor
    Next irow
End If

```

After execution of the rescaling mechanism, as can be seen in **FIGURE 13D**, the height of each row (except the heading rows) has been scaled by *factor*, 0.1 in the example shown. In this manner, all rows of the data are made visible on a single page, thereby facilitating data analysis.

3. Scoring and Sorting Clusters

Scoring and sorting clusters assign numerical scores to the color patterns of individual rows or clusters of rows, thereby enabling comparison and sorting of the clusters by score.

The scoring and sorting mechanism accepts user's designations of colors and corresponding relative scores. It handles cluster numbers or textual labels. Any column can be designated as the cluster labels to be processed. The mechanism scores a user-selected list of columns of data, with user-assigned relative weights, which need not be equal for all columns.

User input is taken from a *Cluster Control* worksheet **314** (see **FIGURES 11A and 11B**), which stores any number of parameter sets, each one with a user-specified name.

The input data is automatically sorted by cluster label before starting, in order to group the clusters together in case the user has previously sorted the data by some other criterion. Then scores are normalized to remove the effects of cluster size, absolute magnitude of scoring points chosen, and absolute size of weights chosen. The results are written to two new worksheets without altering the original data sheet. The first derived sheet is for the numerical scores; the second is like the original, but has the clusters sorted into descending score order,

so that the “best” are at the top, removing the need to visually scan a long colored worksheet. The derived output sheets have names that indicate their source data sheet and the name of the parameter set used for scoring. At the user’s option, the system reversibly hides the un-scored columns in the cluster-sorted output sheet,
5 focusing attention on the data that were used in scoring.

In preferred embodiments, the system detects uncolored cells in the data and offers the user two programmed modes of dealing with them, (uncolored = entry on user’s list of scores or uncolored = “average of other colors in row”), or the option of stopping to color them manually.

10 If the user designates a column of individual compound labels as the “cluster labels,” then the system compares single compounds rather than clusters.

The mechanism operates as follows, with reference to **FIGURES 11A-11C**.

- (1) On the “Cluster Control” sheet **314** the user picks a control panel (e.g., panel **1100**) to use (a panel which is empty or one containing non-
15 needed inputs). On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input information for several tools.
- (2) In the top black section **1102** of the selected control panel **1100**, the user gives this new parameter set a name if not already done. The
20 name will be used to label the outputs.
- (3) In the blue section, the user enters a sheet name (in **1104**) and the column (in **1106**) to be considered as the cluster labels. Note: To score each compound separately rather than in clusters, enter a column with individual compound labels as the “Cluster Col.”

(4) The red section of the control panel is divided into two parts, with its action button **1108**, with red text “Score and Sort Clusters”, in the middle. Above the button, enter the names of the colors **1110** to be assigned point scores, along with their corresponding point scores **1112**. The scores are arbitrary and relative; they will be normalized by the system as necessary. However, a user should be sure always to assign higher point scores to colors which denote favorable values, and lower point scores to colors which denote unfavorable values. The cells with entries need not be colored, and need not be in score order, because the system will color and sort these cells when run.

When assigning point values, a user should be aware that uncolored cells (which are most likely blank, i.e., unknown data) may have quality values above or below those that contain grouped and colored data.. The user may decide that some of the colored groups are “better” or “worse” than data being unknown, and can assign a score to the color “NONE” accordingly.

(5) Below the “Score and Sort Cluster” button **1108**, the user enters the columns **1114** to use for scoring (using the same syntax as for the Data Coloring) and their corresponding relative weights **1116**. The numbers for weights are arbitrary and relative; they will be scaled by the system as necessary. Note that a line with multiple columns will assign the entered weight to *each* of the columns.

(6) The user the selects (clicks) the red-text “Score and Sort Clusters” button **1108**.

(7) When the scoring and sorting tool runs (on the system 100), if the system detects uncolored cells in the data, the user will be offered two modes of dealing with them automatically, or a third manual option of stopping to color them. The two modes are:

- 5 • "Use score for the color "none" on my list"
- (RECOMMENDED)
- "uncolored = average of other colors in row".

(8) The program then scans the chosen columns in each row and adds up
10 the chosen column's color scores for that row. These scores are then
averaged for each cluster of rows, as defined by the user-selected
"cluster column." The VB program code which accomplishes this is as
follows:

```

15     For icol = 1 To ncols
         colorcode =
         Cells(irow, colnum(icol)).Interior.ColorIndex
         colorfound = False
         ' Add up the weighted scores
20     For j = 1 To ncolors
         If (icolor(j) = colorcode) Then
             jscore = j
             colorfound = True
             Exit For
25     End If
         Next j
         If colorcode = xlNone And treatblanks = "AVERAGED"
         Then colorfound = True
         If Not colorfound Then
30         Cells(irow, colnum(icol)).Select
         If colorcode = xlNone Then

```



```

        thiscname = "none"
    Else
        Call ColorNameToIndex(thiscname, colorcode, True)
    End If
5   addscore = score(jscore) * colweight(icol)
    cmpdscore = cmpdscore + addscore
    ' Next IF-THEN-ELSE block is
    ' special calculations for the "averaged" mode
    If colorcode = xlNone Then
10      lostweight = lostweight + colweight(icol)
        minscore = Application.Min(minscore, 0)
        maxscore = Application.Max(maxscore, 0)
    Else
        cmpdscore2 = cmpdscore2 + addscore
15      End If
Next icol

```

(9) The scores are then normalized for the various cluster sizes (number of rows per cluster), and scaled to a value of one hundred (100) for a row which is colored entirely with the user's highest-scoring color and a value of zero for a row that is colored entirely with any color to which the user has assigned a score of zero.

```

If clusterscore(icluster) = 0 Then
25   ' do nothing
ElseIf clusterscore(icluster) > 0 Then
    clusterscore(icluster) =
        100 * clusterscore(icluster) / (nrows * maxscore)
ElseIf clusterscore(icluster) < 0 Then
30   clusterscore(icluster) =
        100 * clusterscore(icluster) /
            (nrows * (-minscore))
End If

```

(10)The results are presented as two newly inserted worksheets. The first is named by appending the word "SCORES" to the name of the original data sheet, and contains a list of the clusters with their sizes and scores.

5 (11)The second new sheet is named by appending the word "SORTED" to the name of the original data sheet. The "SORTED" sheet contains a copy of all the original data and coloring, but with the rows re-ordered to place the highest-scoring clusters at the top, and all the clusters in descending score order from there down.

10 (12)The user has two additional options regarding the appearance of the "SORTED" sheet: (a) a column containing the numerical scores can be added; and (b) the columns that were not used in the scoring can be hidden, so that only the ones actually used remain visible.

15 An example of user provided data is shown in the control panel in **FIGURE 11D** which is taken from the cluster control worksheet shown in **FIGURE 11A**. As shown in **FIGURE 11A**, the parameters are stored with the name "Cmpd" 1102. The scoring a sorting parameters in the control panel 1100 of **FIGURE 11D** give the color red a score of "-1", orange has a score of "0", yellow has a score of "1" and light green has a score of "2". Columns C and D have relative weights of "1", as does column E.

20

Note on the output of score and sort clusters: The system inserts two new sheets after the data (see, e.g., **FIGURES 14C-14D**). The first added sheet contains two score columns: the scores generated by *both* of the auto modes (uncolored = zero and uncolored = average), but the one not selected will be gray. The scores

25

are on a scale of “-100” to “+100”, where a score of “-100” means that all cells had the maximally negative score available, and a score of “+100” means that all cells had the maximally positive score available. The second added sheet has clusters sorted according to the *one* auto mode chosen when the tool ran. The routine offers to hide all columns that were *not* used in the scoring and sorting.
The user can selectively unhide certain columns by using the “Edit:GoTo” menu option (or typing “CTRL-G”), enter the columns in the “Reference” box (for example, C:F), then pick the “Format:Column:Unhide” menu option.

If the user wants to see a color-score-sorted list of compounds within a particular cluster (such as the best cluster), the user should do the following:

1. Sort by clusters to find the ID of the cluster wanted.
2. With a second rule, sort by compounds.
3. Go to the “SORTED by Compound” results sheet and turn on EXCEL’s “Data:Filter:AutoFilter” feature for the column that specified the clustering in the first sort. The user can then choose to view only the compounds in one particular cluster, and they will be in compound-sorted order.

Example

With reference to the already-colored worksheet “DEMO 1” shown in FIGURES 8A-8B, the cluster control worksheet shown in FIGURE 11A, and the control panel shown in FIGURE 11D, application of the scoring and sorting of clusters is described. As noted above, in the control panel of FIGURE 11D, the parameters are stored with the name “Cmpd” 1102. The color red has a score of

“-1”, orange has a score of “0”, yellow has a score of “1” and lightgreen has a score of “2”. Columns C and D have relative weights of “1”, as does column E.

Application of control panel “Cmpd” of **FIGURE 11D**, by selecting “Score and Sort Clusters”, produces the worksheets shown in **FIGURES 14A-14B**. When
5 the user selects the “Score and Sort Clusters” button **1108** for the “Cmpd” control panel of **FIGURE 11D**, the system first presents a dialog box (**1402** shown in **FIGURE 14A**) asking the user how un-colored cells should be scored for sorting. As noted above, un-colored cells can be scored explicitly by user entries (recommended) or as the average of the colors in the same row. Once the user
10 makes a selection and clicks on the “OK” button, the system scores and sorts the data, producing the display screen shown in **FIGURE 14B**. The system provides a summary of what was done, including the information about the two new sheets (“DEMO 1 SCORES by Cmpd” and “DEMO 1 SORTED by Cmpd un=ze”) which are added to the workbook. **FIGURES 14C-14D** show the data in the newly
15 created worksheet “DEMO 1 SCORES by Cmpd”.

Example

With reference to the already-colored worksheet “DEMO 2” shown in **FIGURES 10A-10B**, the cluster control worksheet shown in **FIGURE 11A**, and the control panel shown in **FIGURE 11C**, application of the scoring and sorting of
20 clusters is described. In the control panel of **FIGURE 11C**, the parameters are stored with the name “acids” (**1102**). The color red has a score of “0”, orange has a score of “1”, yellow has a score of “2” and light green has a score of “3”. Column D has a relative weight of “1”.

The application of the parameters or rules in the “acids” control panel produces two new worksheets (“DEMO 2 SORTED by acids” and “DEMO 2 SCORES by acids”) shown in FIGURES 15A-15B.

5 4. Score and Sort Dose-Response Data.

Data grouping and visualized by color coding has also been found to enable an automated solution to another vexing pattern recognition problem. An HTS lab is currently able to provide dose-response data on some subset of the whole collection of compounds originally tested. Sometimes, logistical
10 constraints (time and/or cost) dictate that only a few concentration points can be run on each compound, and the high-throughput nature of the process generates somewhat noisy data. A similar situation sometimes exists in other biological laboratories where assays are very time-consuming. Dose-response curves with few, noisy points are difficult to analyze by traditional curve-fitting methods. The
15 present invention includes a mechanisms/algorithms for analyzing percent-of-maximal-effect data and accurately ordering the compounds by potency, even when faced with few points and high noise.

The mechanism recognizes two properties of the dose-response data for each compound:

- 20 1. “Dose-responsiveness,” the drop-off of activity with dilution, is taken as a sign that the compound has some reasonable pharmacological mechanism of action.

2. The activity measurements at the various concentrations also provide a confirmation of the general level of each compound's activity that was indicated by the original single-poke HTS hit.

These two properties are somewhat independent, as illustrated by the example of a compound that is 95% active at all tested concentrations. It demonstrates very poor (i.e., no) dose-responsiveness over the range of concentrations tested, but is so active that it should not be ignored, because it might reveal a dose response if tested at even lower concentrations.

By using the data groupings and color codes of the dose-dependent activity data columns, which help to smooth out excessively fine distinctions in the numbers, this invention includes an algorithm to assign numerical scores for dose-responsiveness and overall activity in the dose-response data. Moreover, the algorithm also calculates a smart composite of these two scores, in such a way that a highly active compound will get a high composite score even if its dose-responsiveness is poor. This composite score is capable of extracting useful information, even from very noisy data, and has been validated to correctly order a list of test compounds. The system of this invention adds data columns that report all three scores for each compound, and these columns can themselves be color coded, and thus used in further comparison to other types of data by compound or cluster scoring and sorting as described above.

Moreover, within certain limits, the invention's dose-response scoring algorithm can also be used to make *quantitative* estimates of IC_{50} values of compounds, even in the presence of large amounts of experimental error. This is accomplished by adding a set of hypothetical marker compounds with known potencies and theoretically calculated activities at the test concentrations. Since

the ordering algorithm is reliable, these markers will be ordered into their appropriate place, and can be used to calibrate the ordering scores in terms of actual IC_{50} 's. In other words, estimates of IC_{50} for the compounds can be generated by interpolating between the markers in the ordered list of composite scores.

Scoring and sorting dose-response data according to the present invention processes several columns of colored dose-response data (activity vs. concentration) to assign three numerical scores that can later also be color coded, and thus used by the "Score and Sort Clusters" mechanism (described above) to compare compounds or clusters of compounds. The three scores are:

- (a) degree of dose-responsiveness over the concentration range tested;
- (b) overall activity level demonstrated in the dose-response data; and
- (c) a variably weighted composite of (a) and (b), designed to give high scores for high activity even when dose-responsiveness is poor (e.g., a compound that is highly active at all concentrations).

The scoring and sorting dose-response data according to the present invention bases its scoring on colors rather than absolute activity numbers. The mechanism takes user input from a *Cluster Control* worksheet, e.g., as shown in FIGURES 11A-11B. FIGURE 11B shows a control panel from the cluster control worksheet shown in FIGURE 11A, wherein the user has selected columns F to I of worksheet "DEMO 1" for scoring dose-response.

The system detects uncolored data, notifies the user, and asks whether to continue. If yes, the system skips the row containing the uncolored data. The system inserts three new columns on the original spreadsheet to contain the new scores, the new columns immediately following the columns of dose-response data. The column headings show the name of the parameter set used for scoring. Preferably, the system offers to regenerate existing table of Sheet Statistics to correct it for newly added score columns. Further, the system offers to sort the data rows by decreasing score. The system also offers to carry out quantitative estimates of IC_{50} values for the user's compounds, by adding artificial calculated calibration marker compounds.

In order to score and sort dose-response data:

- (1) Ensure that the dose-dependent activity data columns are ordered with highest concentration at the left and lowest concentration at the right. To ensure this, the system will remind the user of this requirement and ask him to confirm it when this tool is run. *Note:* If the data are for an undesired effect such as toxicity, the columns should be ordered the opposite way (lowest concentration left, highest right).
- (2) Use the Data Coloring (described above) to color the dose-response data columns.
- (3) Go to the Cluster Control sheet 314 (FIGURES 11A-11D) and pick a control panel 1100 to use. On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input information for several tools.

- (4) In the top black section, give this new parameter set a name (1102)
if not already done. The name will be used to label the outputs.
- (5) In the blue section, enter a sheet name (1104).
- (6) In the red section (1110), enter the colors used to color the dose-
5 dependent data, and relative point scores (1112) to be assigned to
these colors.
- (7) In the green section (1118), enter the columns which contain the
dose-response data (using the same syntax as for Data Coloring).
- (8) Click the green-text "Score Dose-Response" button (1120).
- 10 (9) If the data are expressed as "percent of maximal effect," the user
can follow the prompts to add calibration markers and make
quantitative estimates of IC_{50} 's.

Note on the output of score and sort dose-response: the system inserts
15 three score columns after the dose-dependent data. The three scores are all scaled
to a 0-100 range, and have meanings as follows:

- (a) degree of dose-responsiveness over the concentration range
tested:
- 100 = smoothly decreasing with dilution, spanning the
20 entire range of color groups;
- 75 = flat dose-response; and
- <75 = even more poorly behaved
- (b) overall activity level demonstrated in the dose-response data
- 100 = highest activity color group at all concentrations.

(c) a variably weighted composite of (a) and (b), designed to give high scores for high activity even when dose-responsiveness is poor (e.g. a compound that is highly active at all concentrations).

5

The Dose-Responsiveness Scoring Algorithm

The data columns are ordered left to right, by decreasing concentration.

The scoring algorithm awards positive score points for each dilution step across the data that actually shows a decrease in the activity data *group* (i.e., the color),

10 and to penalize every step that does not. The algorithm uses the following scoring:

- +1 point (awarded) when a dilution step moves to a lower activity group
- 0 points when a dilution step leaves the activity group unchanged
- -3 points (penalty) when a dilution step moves to a higher activity group

15 The maximum score would go to a compound that shows all the possible color group steps in the right direction, and has no reversals. The minimum score would go to a compound with all the possible reversals, and no correct steps. The program then scales the extremes to 100 and 0, in order to present a consistent interface to the user.

20 The relative magnitudes of the scoring parameters were empirically arrived at by testing “complete sets” of color patterns. This is possible because of the data simplification afforded by the value grouping. If we define the following numerical parameters:

C ≡ number of colors used, i.e., number of data value groups

P \equiv number of points measured, i.e., number of different concentrations
(doses) tested,

then the entire “universe” of possible color patterns includes (C^P) different cases.

For some typical values that might be encountered in real HTS data, this total

5 number of cases is manageable in EXCEL, as shown by TABLE 1, below.

TABLE 1. Total Number of Color Patterns

P = # of conc. Points	C = # of color groups	total number of possible cases
3	3	27
3	4	64
4	3	81
4	4	256
5	3	243
5	4	1024
6	3	729
6	4	4096
7	3	2187
7	4	16384*

* For a spreadsheet with a heading row, this exceeds EXCEL's current limit (for EXCEL 95) by one. This value should not exceed the limit for Excel 97.

10 Scoring was done on several of these complete sets. In each set, the results were sorted by decreasing score and compared to “intuition” for general correctness of ordering of dose-responsiveness, and scanned for cases deemed to be clearly out of order. The (+1, -3) score set was found to produce satisfying ordering, while lesser penalties led to poorly ordered results. More objective tests
15 of ordering (described below) were then used to further validate the algorithm

The case of P=3 and C=3 is presented below in its entirety for illustration. TABLE 2 (FIGURE 16F) shows artificial data and processing for twenty seven (27) hypothetical compounds. The “percent inhibition” columns represent assay “data.” If one defines three groups by breakpoints at 33% and 66%, each cell is

assigned to a data group as shown in the middle set of three columns. Here it is clear that the order of compounds in this table is systematic (111, 112, etc.), to illustrate that the complete set is present. The third set of three columns shows color coding, with the darkest being least active and the lightest being most active.

5 Then the data set was processed by the system to yield dose-responsiveness scores, and the results sorted by this score, giving **TABLE 3 (FIGURE 16G)**, the complete set in order of decreasing dose-responsiveness. **TABLE 3** also shows the intermediate step-scoring and unscaled score points, to aid in following and understanding the algorithm. These points are not displayed
10 by the system itself.

The Overall-Activity Scoring Algorithm

The second property of interest to be extracted from the data is the overall activity level exhibited by each compound. As explained above, this is largely
15 independent of the dose-responsiveness.

The data value groups' ordinal index numbers are used as single-point activity measures instead of the original data numbers. Extra weight is given to activity shown at lower concentrations by the simple algorithm of weighting each data column by its serial position, again ignoring the actual concentration values.
20 The scores are then scaled to the range 0 to 100. The results of this scoring on the same complete set are shown in **TABLE 4 (FIGURE 16H)** which has been re-sorted by decreasing overall activity.

The Composite Scoring Algorithm

Comparison of Tables 3 and 4 (FIGURES 16G & 16H) shows clearly that the compound ordering by dose-responsiveness is quite different from the ordering by overall activity. The user (a chemist) could now color-code the new score columns and use them as independent factors in a larger scoring. However, chemists also want a *single* index of compound quality derived from the dose-dependent data. Moreover, a composite index would further help to alleviate the effects of noise on data interpretation, by incorporating more information into the ordering process. This is an "information-based smoothing" of the data. Therefore, a procedure to calculate a third, "smart composite" score from the other two scores was devised.

The general idea is that when selecting good compounds from dose-response data, compounds showing overall high activity should not be discarded for lack of responsiveness. Therefore, the smart composite score should give more weight to the overall activity when the overall activity is high, but lower weight when it is low. A generalized weighted average is written as

$$\text{composite score} = (\text{activity weight})(\text{activity score}) + (\text{responsiveness weight})(\text{responsiveness score})$$

or, defining corresponding symbols:

$$S_C = (W_A)(S_A) + (W_R)(S_R)$$

If the weights are normalized to sum to unity, then this becomes

$$S_C = (W_A)(S_A) + (1 - W_A)(S_R)$$

The activity weight W_A varies with the activity score S_A in such a way as to achieve the desired result.

The functional form of this variation was the subject of empirical testing. It was decided that the limits would be that W_A would approach 0.5 (activity and responsiveness equally weighted) in the limit of low activity, and that W_A would approach 1.0 (responsiveness ignored) in the limit of high activity. The actual
5 variation was encoded as an exponential increase in order to have rather sharp onset of the activity bias at higher activities:

$$W_A = (C_1) \exp [(k)(S_A)] + C_2$$

The value of the coefficient $k=0.06$, for which the activity bias starts to become substantial around an activity score of eighty (80), was chosen for
10 implementation in a preferred embodiment of this invention, according to empirical results. **FIGURE 16I** shows the variation for a few values of k . **TABLE 5** (**FIGURE 16J**) shows all three scores for the example complete set, now sorted by decreasing composite score.

15 The details of the scoring algorithms were arrived at largely by comparing results to intuitive ordering of the test cases in the complete sets. Because the sets were complete, no really new results can be generated by further test sets. However, one can generate test activity data sets from compounds of known potencies, whose real rank ordering is thus known, in order to see more
20 objectively how well the scoring algorithms rank the results.

To this end, a set of pseudo-ligands was hypothesized, with dissociation constants from a fictitious receptor ranging from nanomolar to millimolar ($pK = 9$ to 3). The set included thirty one (31) compounds, with potencies evenly spaced by 0.2 log units (9.0, 8.8, 8.6, ..., 3.4, 3.2, 3.0). A "pseudo-screen" was created
25 which "tested" binding of these ligands at five concentration points in the usual

range: 10^{-5} M, 3×10^{-6} M, 10^{-6} M, 3×10^{-7} M, and 10^{-7} M. Note that the span of potencies exceeds the span of concentrations tested by two log units on each end, so the test set includes both “very active” and “very inactive” compounds relative to the screening concentrations.

Then artificial binding data were created by calculation as follows.

Assuming a simple binding equilibrium of the ligand to a receptor, the “percent inhibition” at a given ligand concentration is equal to the fraction of receptor sites which are occupied by ligand, given by simple equilibrium equations as

10 $p_{inhib} = 100 \bullet (ligand) / [K + (ligand)]$

For a more realistic simulation, artificial random noise was then added to the calculated numbers. The first experiment reported below used noise randomly distributed over the range of ± 10 inhibition percentage points, and the second with noise up to ± 30 inhibition percentage points. Note that this means ten or
15 thirty percentage points of absolute error, not 10% or 30% of the value.

The artificial data were then color-coded according to the mechanisms of this invention (described above) into four color groups, using the simple breakpoints at 25, 50 and 75 percent inhibition. Note that in assigning these breakpoints, no consideration was given to the actual data values. Then the scoring algorithms of this invention were run, and the compounds sorted by the composite score. Rank order numbers were assigned to the compounds, with 1 being the most potent and 31 the least. In cases of ties in the composite score, equal rank numbers were assigned, with a value equal to the average of the rank numbers spanned by the tied group of compounds (e.g., a tie for 2nd and 3rd

would result in each compound being ranked "2.5"). For each experiment, the final rankings were plotted against the "real" rankings by known potency, to test how well the scoring algorithms ordered the compounds. These plots are shown in **FIGURE 16K** (for noise=10) and **FIGURE 16L** (for noise=30).

5 For the experiment with noise up to 10 inhibition percentage points, shown in **FIGURE 16K**, the ranking of the composite scores is "perfect" (in the sense of having no inversions) over the range of tested concentrations ($pK = 5$ to 7). The pseudo-screen is unable to distinguish the potencies of compounds above or below this range.

10 When the noise is much higher (30 percentage points), the ranking of individual compounds is not as precise, but one can identify three cleanly divided "good-medium-bad" groups, as indicated by the dashed boxes on **FIGURE 16L**. Thus, even with this rather extreme noise level, the invention's scoring still successfully prioritizes the compounds into groups. The range where
15 discrimination is effective is still roughly the range of the test concentrations ($pK = 5$ to 7), but has been reduced somewhat by the higher noise. Note that the ranking within this range (the middle boxed group) is still mostly correct, with only one inversion, even for single compounds.

Quantitative Estimation of Potencies

20 With confidence established that the algorithms provide reliable rankings of compounds by potency, it is possible to proceed to making quantitative estimates. The method uses calibration marker compounds.

To understand this method, it is helpful to realize that the concept is analogous to the quantitative use of SDS polyacrylamide electrophoresis gels to

measure protein molecular weights. The proteins are known to migrate through the gel with speeds directly dependent on molecular weight, but it is difficult to calculate the absolute migration rates for a particular experiment. In dose-response scoring, the compounds are known to be properly ordered, but it is not clear how to calculate a potency (e.g., K_{diss} or IC_{50}) directly from the score.

Protein chemists solve the molecular weight problem by running marker proteins, with known molecular weights, in the same gel, then using their band positions as calibration for the unknowns. Analogously, this invention's quantitative estimation method uses hypothetical marker compounds of known potency to internally calibrate the dose-response composite scores for the user's choice of a coloring rule, then interpolates the potencies of the unknowns.

To create markers, the system asks the user to input the concentrations used for each of the dose-dependent activity data columns. The system then picks a set of calibration concentrations, at intervals of 0.5 log units, to span the tested range. For each of these calibration concentrations, a marker compound is created and added to the user's compound list, and artificial data is calculated for each column, from the same simple equilibrium binding equation used above in the validation study (this time with no "noise"):

$$p_{inhib} = 100 \bullet (\text{ligand}) / [K + (\text{ligand})]$$

The marker data are then colored by the same rule used for the user's compounds, and the scoring and sorting algorithm is re-run.

The result is that the markers are sorted into the list according to their potencies, and the potencies of the other compounds can be estimated by

interpolating between the markers, using the composite dose-response scores. To illustrate, a typical section of a sorted list is shown below in **TABLE 6** (**FIGURE 16M**), using four colors.

Potencies for compounds that fall between two markers are calculated by linear interpolation between the logarithms of the markers. Given the various uncertainties in the data values themselves and in the evaluation process, it was found that linear interpolation between markers spaced at 0.5 log unit intervals was sufficiently precise, and no more complex curve fitting was necessary.

Validation of Quantitative Estimation

Validation of the quantitative estimation method followed a procedure very similar to that used to validate the scoring, and using the same sets of test data with various noise levels. As before, the testing concentrations were from 10^{-5} to 10^{-7} M (negative log from 5 to 7). Marker compounds (no noise) were added with pK's from 4.5 to 7.5, and K_{diss} estimates for the noisy compounds were carried out by the interpolation method. The results are shown below for the cases of 10 and 30 inhibition percentage points of noise.

FIGURES 16N and 16P show that the estimates are clearly quite good within the range of the testing concentrations (pK 5 to 7), but the quality of estimation deteriorates quickly beyond those limits, and algorithm does not reliably distinguish among compounds whose potencies are more than a half log unit beyond the testing range. Therefore, it was decided that presently preferred embodiments would not report any estimated values that fell outside the range of concentrations used in the testing data columns. Thus, in the example in **TABLE 6** (**FIGURE 16M**), the lowest testing concentration was 10^{-7} M (= 0.1 μ M). For the

first compound in TABLE 6, the system has estimated a potency with $pIC_{50} > 7$, but it conservatively only reports " $<0.1 \mu M$."

TABLE 7 summarizes the statistics of the estimations within the testing limits. TABLE 7 shows that the method successfully estimates the potencies within about a factor of two, even with high noise levels.

TABLE 7. STATISTICS OF ESTIMATION VALIDATIONS

<i>percentage points of noise</i>	<i>number of compounds</i>	<i>average of abs(log error)</i>	<i>Average error in IC_{50} (factor)</i>
10	13	0.22	x 1.6
30	13	0.39	x 2.5

Comparison to Other Methods of Quantitative Estimation

Further corroboration was obtained by treating some real data from T-cell proliferation blockage assays. It is estimated that these data have *at least* as much noise as the artificial test set with 30 inhibition percentage points added. The standard treatment of this data in the past has been to fit a dose-response curve with a Hill coefficient of 1, using a PC-based program ORIGIN. (ORIGIN is a data analysis program from Microcal Software, Inc. of Northampton, Massachusetts. ORIGIN is used in this instance for non-linear least-squares fitting of dose-response curves to functional equations.)

The data used here were from testing in the concentration range from 1 to $0.03 \mu M$ (negative log from 6 to 7.5). The plot in FIGURE 16Q shows the correlation of values estimated by this invention with values from ORIGIN fits.

Two compounds that the present invention estimates to be beyond the testing range, i.e., pIC_{50} below 6, are included as open diamonds, for illustrative purposes explained below. (As explained above, preferred embodiments of this invention would normally not report these values.)

5 The results are consistent with the properties observed in the validation study. The present invention estimates are quite good within the testing range (6 to 7.5). At the lower limit, this invention has made two estimations exactly at 10^{-6} M (arrows) which do not correlate as well with the ORIGIN fits. Nevertheless, because the whole plot spans only a relatively narrow range of potency, even these
10 discrepancies are not very large. For all twenty eight (28) estimates within the testing range (including these two), the average logarithmic deviation between the two estimation methods is 0.18, corresponding to a factor of only 1.5.

 It is further noted that the calibration marker estimation method does not uniformly “flatten out” beyond the testing concentration range. The two open
15 diamonds in **FIGURE 16Q** are estimations that presently preferred embodiments of this invention would not normally report because they have $pIC_{50} < 6$, but they agree well with ORIGIN fits

 It is interesting to compare calibration-marker with curve-fitting results for particularly badly behaved data, such as dose-response curves that are not
20 monotonic with respect to concentration. This is sometimes the type of data that emerges from dose-dependent screening in a high-throughput mode. Studies of this type have been initiated by adding artificial noise to the extent of fifty (50) inhibition percentage points.

 Finally, it should be pointed out that there is some mechanical advantage
25 to using the present invention relative to current practice of using ORIGIN.

ORIGIN is used by manually filling in a template with data, then manually executing a fit. Depending on the number of points and the degree of customization of parameters, this can take one to ten minutes of the user's time. The present invention, on the other hand, processes a whole spreadsheet at once
5 (i.e., up to 16,383 compounds), and goes at a rate of about 3,000-4,000 compounds per minute on a 200 MHz PC.

Examples

FIGURE 16A shows dose response data for twenty (20) compounds at four
10 concentrations. The data have been grouped and the cells colored by the rule shown in **FIGURE 16B**. The result of the scoring and sorting process is shown in **FIGURE 16C**, where the compounds are ordered by decreasing values of the composite score (column H). Then, virtual "marker" compounds are added with known potencies spaced by 0.5 log units, and they are shown in **FIGURE 16D**,
15 colored by the same rule and scored. The name of each marker compound designates the logarithm of its potency, e.g., "marker_7.0" has a potency $IC_{50} = 10^{-7}$ M. **FIGURE 16E** shows the result of sorting the list by decreasing composite score after adding the markers. This process then enables estimation of IC_{50} values for the compounds by interpolating in the column (H) of ordered composite
20 scores, and these estimates appear in two forms in columns I and J.

6. Summarize Spreadsheet Statistics Mechanism

This mechanism creates a table summarizing the entries in each column of a data sheet, to aid the user in deciding how to color each column. The

mechanism counts numeric, text, and data entries, and uses color to flag columns that have mixed types. The mechanism also counts blanks, and specially flags columns with "trailing blanks," i.e., columns which are shorter than the longest one on the spreadsheet. For numeric data, the mechanism calculates minimum, maximum, mean, and standard deviation, even in the presence of interspersed text entries. For text data, the mechanism presents a list of the text strings used and their occurrence counts. The mechanism creates a summary key of the column letters and headings as a text box that can be copied to other sheets for convenient reference.

FIGURE 17A shows a sample spreadsheet containing miscellaneous data on twenty four (24) compounds. **FIGURE 17B** is the statistics sheet calculated from it. Each row of the statistics sheet describes one column of the original data sheet. First, the counts of numeric, text, date, and blank entries are listed, followed by two columns describing the total length of the data sheet. Then the minimum, maximum, mean, and standard deviation of any numeric data are reported. Finally, the statistics sheet lists a summary of the text strings found in each original data column. As examples, in **FIGURE 17B**, one can see that original column A ("Cmpd") had twenty four (24) different text strings, that the numeric data in original column C ("Test1") had a mean of 2.385, and, flagged by the red coloring, that original data column E ("Test3") had a mixture of ten numeric data and two text strings, both "N.A."

The details of how the program code accomplishes this are straightforward, and one of ordinary skill in the art would know, from this description, including the Figures, how to make and use this invention. The

program loops through all the entries in the column, testing the data type of each, and tallying the counts and numerical statistics.

Spreadsheet Creation and Organization

5 The operations of this invention require a considerable amount of user input, e.g., to create well-structured spreadsheets, to define and apply diverse coloring rules for large numbers of columns, and to use these colors and the user's stated scientific priorities to create meaningfully ordered lists of compounds or clusters.

10 The user interface of this invention has been designed to ease this process and help the scientist focus on the tasks of formulating and recording clear descriptions of the evaluation parameters. Accordingly, this invention provides a number of tools and mechanisms to aid in the creation and organization of spreadsheets. These tools and mechanisms include:

- 15 • Smart Append Column Mechanism
- Merge Data Mechanism
- Data Import Mechanism
- Workbook Navigation Shortcuts
- Conversion of "uM" to μ M and "UU" to μ
- 20 • Delete Pictures Mechanism
- Change Values in Column Mechanism
- Concatenate Values across Columns Mechanism
- Delete Leading Inequality Signs Mechanism
- Delete Derived Sheets Mechanism

Smart Append Column

This mechanism appends new columns of data onto an existing spreadsheet, matching rows by labels (e.g., compound numbers). The mechanism
5 copies all data to a new sheet before doing its work, leaving the original sheets unchanged. There is no need for the user to pre-sort any of the data. The mechanism provides optional case-sensitive or case-insensitive label matching.

New rows are added at the bottom when new labels do not match any old labels. Rows with missing labels are identified and the system offers to fill them
10 by copying previous label. Rows with repeated labels (i.e., replicate data) are also identified and the system offers a choice from among several automated processing rules, or manual fixing. A fast matching algorithm temporarily sorts rows by label, then restores original order when finished. Several intermediate stopping points are offered and extra data viewing options for conservative users
15 worried about errors.

Merge Data Mechanism

The merge data mechanism copies new data values from an appended
20 column into an older column. The mechanism copies all data to a new sheet before doing its work, leaving the original sheets unchanged. The mechanism detects cells where new data would overwrite old data that is different, flags them with color, and alerts the user. Several intermediate stopping points are offered to

the user, as are extra data viewing options are offered for conservative users worried about errors.

Data Import

5 One-button (or one-menu-click) import of existing EXCEL spreadsheets into an integrated file, which contains both the data and the related control sheets. The mechanism offers to search for and remove any leading or trailing spaces in the imported data and offers to consolidate replicate data rows into unique ones, using user choices as to how to handle the replicate data. The mechanism also
10 detects hidden rows and offers to unhide them and detects formulas and offers to convert them to values. This mechanism is also used to update to newer version of the system.

Workbook Navigation Shortcuts

15 The system includes various workbook navigation shortcuts including:

- A special added drop-down menu which includes commands for jumping directly to the various control sheets. These commands also have keyboard shortcuts assigned to them.
- From a cell on a control sheet that contains the name of a data sheet, a
20 special item on the right-click shortcut menu jumps directly to that data sheet. Other special items on this menu enable a “Twin Screen” display to see two sheets at once.

- To aid in choosing columns to enter on control sheets, there is a special “Twin Screen” display triggered by right-clicking any cell on a control sheet that contains the name of a data sheet.

5 Convert “uM” to μ M and “UU” to μ

Preferred embodiments of the system of this invention require the data spreadsheet to have *one and only one* row of column headings. The user can type either of the encoded strings “uM” (lowercase u, uppercase M) or “UU” (both uppercase) into any column heading, select the cell or whole row of headings, then
10 pick this command. Each “uM” in the selection will be converted to “ μ M”, and each “UU” will be converted to a “ μ ”. The code recognizes the special exception of the word “VACUUM” as long as it doesn’t end with the cases “uM.” This conversion allows the user to avoid the confusing use of lowercase “u” or the column-widening use of the full prefix “micro.” This utility appears on the
15 system menu.

Delete Pictures

The system provides a mechanism for removing pictures containing chemical structures, in order to reduce file size, processing time, and confusion
20 when they do not align properly after row sorting.

Change Values in Column

This is a mechanism for regularizing data in a spreadsheet column. It facilitates replacement of all occurrences of a given value by another. The mechanism creates backup copies of the original column, and updates any existing data statistics for the edited sheet.

Concatenate Values across Columns

The system provides a mechanism for regularizing data in a spreadsheet column. Some possible uses include: (a) construction of unique row labels: M-number plus stroke number → "M123456/001"; and (b) reconstitution of numerical inequalities from separate columns: ">" plus a number → ">number". The user is provided with an option to include linking (delimiting) text strings between values and an option to include or skip blanks. The system retains the original columns and inserts a new one for the results.

Delete Leading Inequality Signs

Another mechanism for regularizing data in a spreadsheet column includes the mechanism to delete leading inequality signs. This mechanism converts entries like ">1000" to just the number "1000". This must be used with considerable caution, because it is the equivalent of creating a false test result. It is generally preferable to color the cells containing text strings with the data coloring mechanism described above, rather than alter them. All later processing is based on the colors, not the cell values. This mechanism also deletes inequality

sign only if it is the first character in the cell. The mechanism creates backup copies of the original columns.

Delete Derived Sheets

5 The system menu includes a command to delete all output sheets from the current workbook, with separate user confirmation for each one. This is intended as a cleanup mechanism for information that may be outdated and is easily regenerated by subsequent system runs.

10 Initial experience with the coloring tool has revealed that color coding has more subtle, but far-reaching usefulness. The colors themselves also can act as a “currency of exchange,” a medium for comparing the quality of one kind of result to the quality of a very different kind of result. For example, an HTS activity of “95% inhibition” may be considered desirable and color coded, e.g., green. In the
15 same list of compounds, a molecular weight between 400 and 600 may be considered optimally desirable, and thus also color-coded green. If the user takes care when assigning colors, “green” takes on a common meaning across the board. This translation of data values into colors then opens up a cornucopia of possibilities for processing the colors (as numerical color indices) and comparing
20 compounds, searching, in our example, for the ones that are the “most green.”

 Accordingly, the system includes tools to numerically score individual compounds or clusters of compounds by the colors that appear in their various data columns. The system can then create a new spreadsheet sorted by this score

(either by single compounds, or cluster-by-cluster, the choice being the user's), in which the "most green" compounds will then appear at the top.

5 **Examples**

Application to Portfolio Management

The system can equally well be applied to any set of data where the rows are cases of a similar construct, with the columns being various properties of each case. For example, a data spreadsheet can contain a list of competing projects or
10 investments for a company's portfolio, with the columns containing various managers' ratings of each project or investment. **FIGURE 18A** shows an example of twenty projects, each of which has been scored 1, 2, or 3 on two factors, one more important than the other, by each of three managers. The sheet has been colored by the rule shown in **FIGURE 18B**. Then the data were scored and sorted
15 by the sorting rule of **FIGURE 18C**, and the result is shown in **FIGURE 18D**. Clearly, the projects that were given a "3" in the important factor come to the top, and it can be seen that the less important factor does indeed matter less to the final ordering. The colors also help to flag anomalies, such as a low score by one manager on an otherwise high-ranking project.

20 In general, the data can be various sorts of data. Some examples are listed below and illustrated in the referenced Figures.

FIGURE 19 shows a list of drug candidate compounds, scored and sorted by a composite of ten parameters that describe their physical, chemical, and biological properties. Green shades indicated desirable values; red shades are

undesirable. The display is compressed vertically with the vertical re-scaling tool to clearly display the difference in coloring patterns between the top eighty (80) compounds and the bottom eighty (80) compounds (separated in the illustration by a blank band).

5 **FIGURE 20** shows a list of proteins that are candidates for targets for drugs, chosen from a pool of candidate genes, scored and sorted by a composite of eleven parameters that describe their suitability.

FIGURE 21 shows a list of research projects competing for resources. Each project has been scored according to several evaluation factors, and the whole
10 array has been sorted by color groups. The same construct is useful for evaluating employee performance or job candidates.

FIGURE 22 shows a list of pharmaceutical companies and their current status with regard to discovering or marketing products in each of various disease areas. Each company's line of products has been scored according to the maturity
15 of the offerings, and the whole array has been sorted by color groups.

FIGURE 23 shows the use of data-grouping (coloring) rules to visualize the time courses of drug concentrations in blood. In this example, light colors were chosen to represent high concentrations of drug in the blood, while dark colors were chosen to represent low concentrations. The figure shows a wide range of
20 differing time courses.

FIGURE 24 shows the use of data-grouping (coloring) rules to visualize the matrix of pairwise cross-correlations of the results of eight (80) drug screens. In this example, light colors were chosen to represent low correlations, while dark colors represent high correlations.

Quantitation of the Similarity of Data Grouping in Two Variables

As part of the present invention, a mechanism is provided for assigning a quantitative measure to the degree of similarity of grouping (visualized by color coding) of data in each of two columns of an EXCEL spreadsheet. The mechanism
5 allows for a correlation-like analysis on a wide variety of data types, including text, or mixed numbers and text.

In the data-exploration paradigm of the present invention, one of the first steps a user takes is to divide the range of data values in each column into a small number of groups for further analysis, thus effecting a reduction of precision
10 which has been found to be useful in various ways.

It is sometimes useful to explore whether the rows of the data matrix have been divided into similar groupings in each of two different columns. For example, a researcher might ask, "Do the high molecular weight compounds tend to be the ones whose solubilities fall below the limits of measurement?" In other
15 words, this would mean to compare the groupings in the molecular weight column with the groupings in the solubility column.

If the data were strictly quantitative, this would be called *correlation* of variables, and there exist a number of perfectly good statistical measures of the phenomenon. However, one of the unique capabilities of the present invention
20 lies in dealing with textual data and mixtures of numbers and text, and it would be helpful if one could translate the visible color patterns of the present invention to some kind of quantitative measure of correlation. In order to avoid confusion with standard statistical correlation, the distinct term "color grouping similarity" is used to describe the new measure.

In preferred embodiments of this invention, the data grouping is stored in the form of the colored backgrounds of data cells. At first glance, one might simply seek to compare the colors of the first column with those of the second, and count the number of rows with matching colors. However, a color grouping
5 similarity tool must be able to cope with the possibility that the colors are different. This could happen because the user chooses completely *different color schemes* for the two columns, or because the correlation is negative. As an example of negative correlation, suppose column A contains random numbers between 0 and 1, colored such that those above 0.5 are green and those below 0.5
10 are red. Then imagine a column B where each value is equal to one minus the corresponding value in column A, i.e., the “one’s complement.” If the user colors the second column with exactly the same coloring rule as the first, every row will have a different color in column B than in column A. None of the colors will match, though the groupings are perfectly correlated. To be successful, the tool
15 must deliver a high measure of correlation between such pairs of columns. The algorithm described below was designed to perform in this way.

Algorithm for Measuring Data Grouping Similarity

The algorithm was derived from semi-quantitative reasoning, as follows.
20 It is based on the qualitative question, “For all rows that have one particular color in the first column, to what degree do they have a *uniform* color in the second column (not necessarily the *same* color as in the first column)?” The quantified answer to this question is then averaged over the set of colors used.

The details of the mechanism can be seen by example. First, to compare the grouping in two columns A and B, a matrix of “ordered color pair counts” (OCPC) is defined such that each matrix element $OCPC_{ij}$ is the count of spreadsheet rows where one finds color i in spreadsheet column A and color j in spreadsheet column B. Then, the rows of the two spreadsheet columns are scanned to count the number of occurrences of each ordered color pair and thus to determine the values of the matrix elements.

In the discussion below, carefully distinguish the rows and columns of the user’s *spreadsheet* of data from the rows and columns of the derived OCPC *matrix*.

If the two spreadsheet columns had *exactly the same coloring*, the nonzero elements of the OCPC matrix would all be on the diagonal. As a simplified example, consider four colors (green, yellow, orange, red) and a total of 16 data rows. The diagonal matrix might be (zero elements left blank for emphasis)

color in column A	color in column B →	green	yellow	orange	red
green		4			
yellow			2		
orange				7	
red					3

In the case above, there are groups of 4 spreadsheet rows colored green (in both columns), 2 rows colored yellow, 7 rows colored orange, and 3 rows colored red.

In contrast, if the groupings were the same, but the coloring rule for the second spreadsheet column used the same colors in a different order, the OCPC matrix might look like the following, no longer diagonal:

color in column A	color in column B →	green	yellow	orange	red
green				4	
yellow		2			
orange					7
red			3		

A simple extension applies if *entirely different colors* (cyan, blue, maroon, purple) are used in the second spreadsheet column. The OCPC matrix might then be:

color in column A	color in column B →	cyan	blue	maroon	purple
green				4	
yellow		2			
orange					7
red			3		

In *any* of the three cases above, the groupings are identical, and the OCPC matrices have the property that each matrix row and matrix column has only one nonzero element. That lone element is necessarily equal to the sum of the row or column. This situation should receive the *highest* similarity score.

One way to define the contrasting situation that would deserves the *lowest* similarity score would be that for each user-defined group of spreadsheet rows in one spreadsheet column, the other spreadsheet column has a “maximally non-unique” set of colors. In the corresponding OCPC matrix, this corresponds to each matrix row or column having a broad distribution of values rather than a single non-zero, a uniform distribution has been chosen as the definition of this state:

color in column A	color in column B →	cyan	blue	maroon	purple
green		1	1	1	1
yellow		1	1	1	1
orange		1	1	1	1
red		1	1	1	1

This low-similarity state can be more precisely defined by saying that each element in a given matrix row or column is the average of all the counts in that matrix row or column.

5 With these concepts defined and the OCPC matrix filled, the scores can then be derived. Each OCPC matrix row (corresponding to a color group in the first spreadsheet column) is selected in turn for scoring. Each element in the matrix row is given a score between zero and one, according to its linear interpolation between: on the one extreme, the average of the nonzero elements in
10 the row, and on the other, the sum of the row or column (i.e., the maximum value it could have if all the others were zero). The scores are then averaged over all the rows of the OCPC matrix to generate a row-wise score component.

 Next, the corresponding process is applied to the *columns* of the OCPC matrix (each corresponding to a color group in the second spreadsheet column
15 rather than the first). The resulting column-wise score component is averaged with the row-wise score component, then the average is scaled to a maximum of 100 to generate the final similarity score for the two spreadsheet columns.

Interpretation of the Similarity Scores

20 Although the scores are quantitative and well-defined, their interpretation is best done in a partly subjective manner, based on experience. The behavior of

the scores is best understood by example. In FIGURE 25, the leftmost (“base”) column has been compared to each of the others, and the scores are shown as well as pictures of the grouping patterns. Comparison of the base with the next column shows that the tool delivers a maximal score of 100 for identical grouping, even when the colors are completely different. Then, stepping across the figure toward the right, it can be seen how the score decreases as the grouping pattern gradually becomes less similar to that of the base column. All the way down to a similarity score of 40, it is still basically true that the light colors are on top and the dark on the bottom, with increasing “noise,” but when the score falls to 20, the pattern appears to have no correspondence to that of the base.

Implementation of the Data Grouping Similarity Tool

In practice, the tool allows the user to choose two sets of spreadsheet data columns. The program then automatically generates all pairs containing a column from the first set with a column from the second set, then writes the similarity scores onto a newly inserted spreadsheet in the user’s workbook. The output takes the form of a table where the degree of similarity is itself color-coded to aid the user in identifying significant cases. An example appears FIGURE 26.

While the invention has been described with reference to particular mechanisms (algorithms, processes and functions) and architectures, one skilled in

the art would realize that other mechanisms and/or architectures could be used while still achieving the invention.

While embodiments of the present invention have been described with particular setup and initialization procedures, other setup and/or initialization
5 procedures can be used.

Further, while many of the operations have been shown as being performed in a particular order, one skilled in the art would realize that other orders, including some parallelization of operations, are possible and are considered to be within the scope of the invention.

10 While the present invention has been described with reference to analysis and pattern recognition in data sets relating to chemical compounds, the methods, systems and devices of this invention are considered to be general constructs covering other, non-chemical data sets.

Thus, are provided methods, systems and devices for analysis and pattern
15 recognition in large, multidimensional data sets using low-resolution data grouping. One skilled in the art will appreciate that the present invention can be practiced by other than the described embodiments, which are presented for purposes of illustration and not limitation, and the present invention is limited only by the claims that follow.

What is claimed is:

1. A method of operating on data, the method comprising:
providing at least one user-defined grouping rule for grouping the data into
5 a user-definable number of groups; and
applying at least one of the grouping rules to the data.
2. A method as in claim 1 wherein the data are provided in a table and
wherein the at least one grouping rule applies to at least one user-selectable
10 column of the table.
3. A method as in claim 1 wherein the at least one grouping rule
defines breakpoints corresponding to the user-definable number of groups, and
wherein application of the at least one rule to the data divides the data into groups
15 based on the breakpoints.
4. A method as in claim 1 further comprising:
presenting the grouped data in a manner that visually distinguishes the
groups.
20
5. A method as in claim 4 wherein the grouping rules associate colors
with groups and wherein the presenting of the grouped data further comprises:
coloring an aspect of the data according to the rules.

6. A method as in claim 4, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein the presenting of the grouped data comprises:

5 coloring each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

7. A method as in any one of claims 3 and 6, wherein the breakpoints
10 are selected from: (a) numeric values; and (b) textual values.

8. A method as in claim 3 wherein the at least one breakpoint is determined automatically based on the data.

9. A method as in claim 5 wherein the data are provided in a table,
15 wherein the coloring of an aspect of the data comprises:
coloring backgrounds of table cells according to the rules.

10. A method as in claim 1 wherein the number of groups is fewer than
20 a number of possible data values.

11. A method of operating on data, the method comprising:
providing at least one user-defined grouping rule for grouping the data into
a user-definable number of groups;

applying at least one of the grouping rules to the data to generate grouped data;

providing at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and

5 applying at least one of the scoring rules to the grouped data to score the grouped data.

12. A method of operating on data, the method comprising:

generating grouped data by applying to the data at least one user-defined
10 grouping rule for grouping the data into a user-definable number of groups; and
scoring the grouped data by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

13. A method according to claim 11 or 12 wherein the data comprises a
15 number of parameters for each of a number of cases and the scoring rule
comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the scoring of the grouped data comprises:

applying the function to the data to obtain a score for each case.
20

14. A method according to claim 13, further comprising:
sorting the scored cases by score.

15. A method according to claim 14, wherein the scored cases are
25 sorted individually.

16. A method according to claim 14, wherein the scored cases are sorted by cluster.

5 17. A system for operating on data, the system comprising:
a mechanism constructed and adapted to provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and
a mechanism constructed and adapted to apply at least one of the grouping rules to the data.

10

18. A system as in claim 17 wherein the data are provided in a table and wherein the at least one grouping rule applies to at least one user-selectable column of the table.

15

19. A system as in claim 17, wherein the at least one grouping rule defines breakpoints corresponding to the user-definable number of groups, and wherein application of the at least one rule to the data divides the data into groups based on the breakpoints.

20

20. A system as in claim 17, further comprising:
a mechanism constructed and adapted to present the grouped data in a manner that visually distinguishes the groups.

21. A system as in claim 20, wherein the grouping rules associate colors with groups and wherein the mechanism constructed and adapted to present the grouped data further comprises:

5 a mechanism constructed and adapted to color an aspect of the data according to the rules.

22. A system as in claim 20, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein
10 the mechanism constructed and adapted to present the grouped data comprises:

a mechanism constructed and adapted to color each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

15 23. A system as in any one of claims 19 and 22, wherein the breakpoints are selected from: (a) numeric values; and (b) textual values.

24. A system as in claim 19 further comprising:

a mechanism constructed and adapted to determine at least one breakpoint
20 automatically, based on the data.

25. A system as in claim 21 wherein the data are provided in a table, wherein the mechanism constructed and adapted to color an aspect of the data comprises:

a mechanism constructed and adapted to color backgrounds of table cells according to the rules.

26. A system as in claim 17 wherein the number of groups is fewer
5 than a number of possible data values.

27. A system of operating on data, the system comprising:
a mechanism constructed and adapted to provide at least one user-defined
grouping rule for grouping the data into a user-definable number of groups;
10 a mechanism constructed and adapted to apply at least one of the grouping
rules to the data to generate grouped data;
a mechanism constructed and adapted to provide at least one user-defined
scoring rule for scoring the grouped data according to user-defined scores; and
a mechanism constructed and adapted to apply at least one of the scoring
15 rules to the grouped data to score the grouped data.

28. A system of operating on data, the system comprising:
a mechanism constructed and adapted to generate grouped data by
applying to the data at least one user-defined grouping rule for grouping the data
20 into a user-definable number of groups; and
a mechanism constructed and adapted to score the grouped data by
applying to the grouped data at least one user-defined scoring rule for scoring the
grouped data according to user-defined scores.

29. A system according to claim 27 or 28 wherein the data comprises a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the mechanism constructed and adapted to score of the grouped data comprises:

a mechanism constructed and adapted to apply the function to the data to obtain a score for each case.

30. A system according to claim 29, further comprising:
a mechanism constructed and adapted to sort the scored cases by score.

31. A system according to claim 30, wherein the scored cases are sorted individually.

32. A system according to claim 30, wherein the scored cases are sorted by cluster.

33. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:

providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and

applying at least one of the grouping rules to the data.

34. A medium as in claim 33 wherein the data are provided in a table and wherein the at least one grouping rule applies to at least one user-selectable column of the table.

5 35. A medium as in claim 33 wherein the at least one grouping rule defines breakpoints corresponding to the user-definable number of groups, and wherein application of the at least one rule to the data divides the data into groups based on the breakpoints.

10 36. A medium as in claim 33, wherein the method further comprises: presenting the grouped data in a manner that visually distinguishes the groups.

15 37. A medium as in claim 36 wherein the grouping rules associate colors with groups and wherein the presenting of the grouped data further comprises: coloring an aspect of the data according to the rules.

20 38. A medium as in claim 36, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein the presenting of the grouped data comprises: coloring each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one
25 breakpoint.

39. A medium as in any one of claims 35 and 38, wherein the breakpoints are selected from: (a) numeric values; and (b) textual values.

5 40. A medium as in claim 35 wherein the at least one breakpoint is determined automatically based on the data.

41. A medium as in claim 37 wherein the data are provided in a table, wherein the coloring of an aspect of the data comprises:
10 coloring backgrounds of table cells according to the rules.

42. A medium as in claim 33 wherein the number of groups is fewer than a number of possible data values.

15 43. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:

providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups;

20 applying at least one of the grouping rules to the data to generate grouped data;

providing at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and

25 applying at least one of the scoring rules to the grouped data to score the grouped data.

44. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:

5 generating grouped data by applying to the data at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and
 scoring the grouped data by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

10 45. A medium according to claim 43 or 44, wherein the data comprises a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the scoring of the grouped data comprises:

15 applying the function to the data to obtain a score for each case.

46. A medium according to claim 44, the method further comprising:
 sorting the scored cases by score.

20 47. A medium according to claim 46, wherein the scored cases are sorted individually.

48. A medium according to claim 46, wherein the scored cases are sorted by cluster.

25

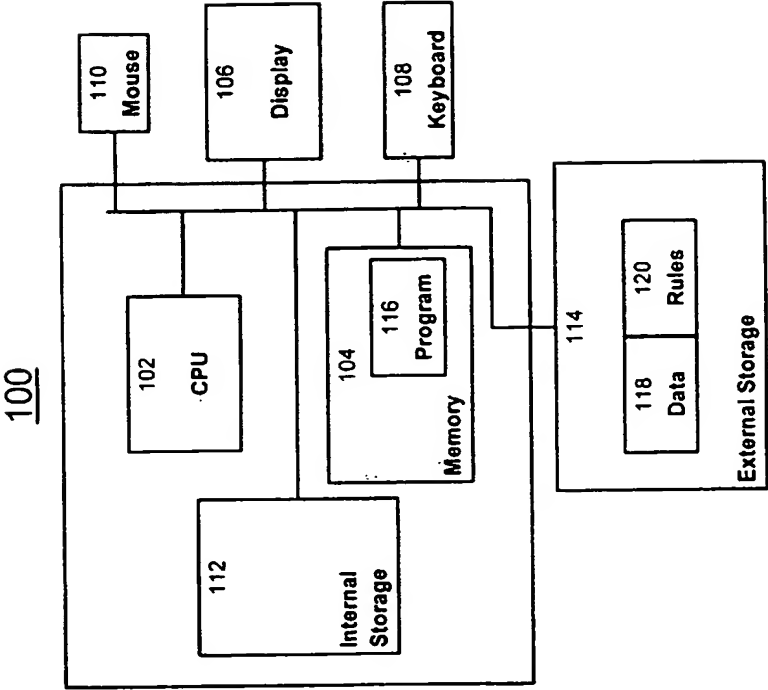
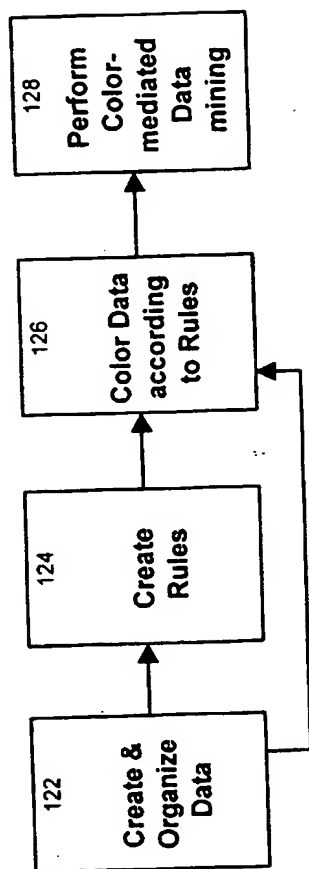


Fig. 1

Fig. 2

300

Microsoft Excel - demo - PANDORA v23.1.xls [Read-Only]										
A1 Cmpd										
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M		
Cmpd01	N		29	30	41	3	22	5		
Cmpd02	N		42	55	83	57	28	15		
Cmpd03	G		261	11						
Cmpd04	N			30	70	25	24	29		
Cmpd05	N		18	32	89	60	21	22		
Cmpd06	D	886	6.5	37	71	41	13	3		
Cmpd07	D	311	0.037	78	100	79	48	43		
Cmpd08	D		0.089	26	65	28	28	38		
Cmpd09	D	0.119			68	41	22	15		
Cmpd10	N	0.233			61	42	24	5		
Cmpd11	N	4.31			50	77	63	25		
Cmpd12	H	1.3	0.24		47	25	40	37		
Cmpd13	H	1.17	0.194	30	81	59	4	12		
Cmpd14	H	0.26	0.41		39	23	4	36		
Cmpd15	H	0.369	0.148		99	46	38	18		
Cmpd16	I		0.87	30	101	82	47	24		
Cmpd17	K		0.223	30	81	64	22	32		
Cmpd18	I	527			79	54	23	12		
Cmpd19	I	0.134			71	71	108	100		
Cmpd20	F		0.317		101	109	31	13		
Cmpd21	K		2.21		87	70	35	12		
Cmpd22	B		0.15		94	77	36	12		
Cmpd23	B				96	61	69	39		
Cmpd24	B		0.27		110	91	75	52		
Cmpd25	B	0.041	1.1		105	104	41	22		
Cmpd26	B	0.665			93	71	43	23		
Cmpd27	B	0.111			97	79	52	25		
Cmpd28	E		0.13		95	62	12	11		
Cmpd29	J		0.46		68	48	9	17		
Cmpd30	J	2.79	45	52	41					
Cmpd31					DEMO 1					

Fig. 3A

300

File Edit View Insert Format Tools Data Window Help											
Ready											
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M	HTS SPA Dose-Resp % Inhib @ 3x10-8M	HTS SPA Dose-Resp % Inhib @ 1x10-8M	HTS SPA Dose-Resp % Inhib @ 3x10-9M
Cmpd34	J	2.27	24	30	105	62	62	21			
Cmpd35	J	0.63			41	28	18	3			
Cmpd36	J		0.23		93	63	42	14			
Cmpd37	N	12.3			58	20	38	30			
Cmpd38	L	0.009	0.024		90	60	61	2			
Cmpd39	F	0.55	0.41		87	73	29	15			
Cmpd40	G	0.359	1.04		65	65	40	11			
Cmpd41	G	0.018	0.35		66	46	28	13			
Cmpd42	F	0.39	0.9		102	97	92	97			
Cmpd43	H	4.45	0.13		36	25	18	12			
Cmpd44	G	0.076	0.1		78	60	40	25			
Cmpd45	F	0.5			111	110	104	82			
Cmpd46	M	1.14			25	21	16	21			
Cmpd47	F	0.026			109	104	97	80			
Cmpd48	F	0.27			100	102	83	75			
Cmpd49	M	0.035	0.043		71	43	31	15			
Cmpd50	F	0.051			112	111	112	75			
Cmpd51	G	0.079			78	70	44	27			
Cmpd52	H	0.33	0.117	4.9	69	24	23	14			
Cmpd53	A	0.035	0.18	1.6	108	102	63	27			
Cmpd54	A	0.33			91	78	75	33			
Cmpd55	G	9.12			61	64	60	28			
Cmpd56	C	0.39			93	80	55	37			
Cmpd57	A	0.018			101	72	42	29			
Cmpd58	C	0.22			92	69	55	49			
Cmpd59											
Cmpd60											
Cmpd61											
Cmpd62											
Cmpd63											
Cmpd64											
Cmpd65											
Cmpd66											
Cmpd67											
Cmpd68											
Cmpd69											
Cmpd70											
Cmpd71											
Cmpd72											
Cmpd73											
Cmpd74											
Cmpd75											
Cmpd76											
Cmpd77											
Cmpd78											
Cmpd79											
Cmpd80											
Cmpd81											
Cmpd82											
Cmpd83											
Cmpd84											
Cmpd85											
Cmpd86											
Cmpd87											
Cmpd88											
Cmpd89											
Cmpd90											
Cmpd91											
Cmpd92											
Cmpd93											
Cmpd94											
Cmpd95											
Cmpd96											
Cmpd97											
Cmpd98											
Cmpd99											
Cmpd100											

Fig. 3B

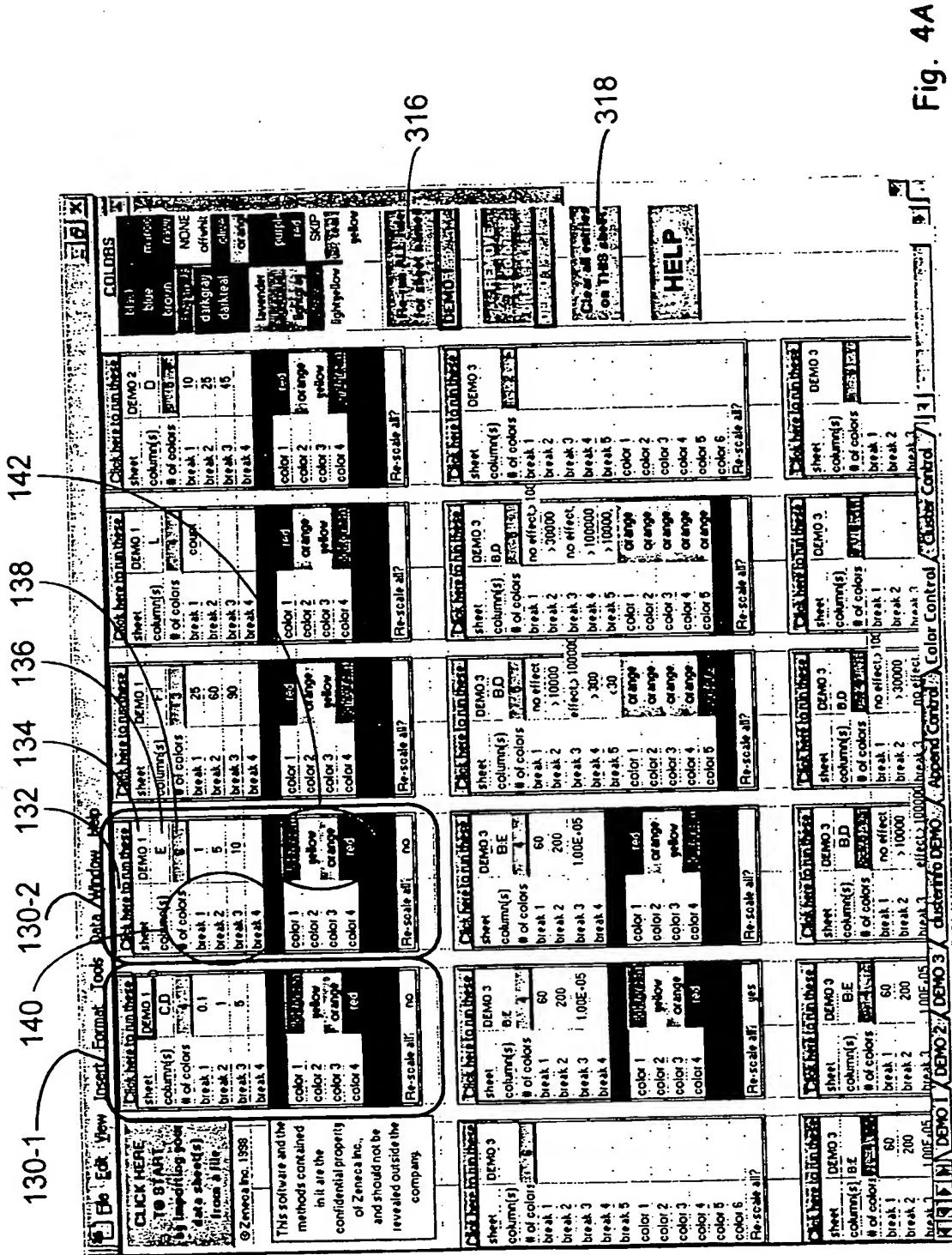
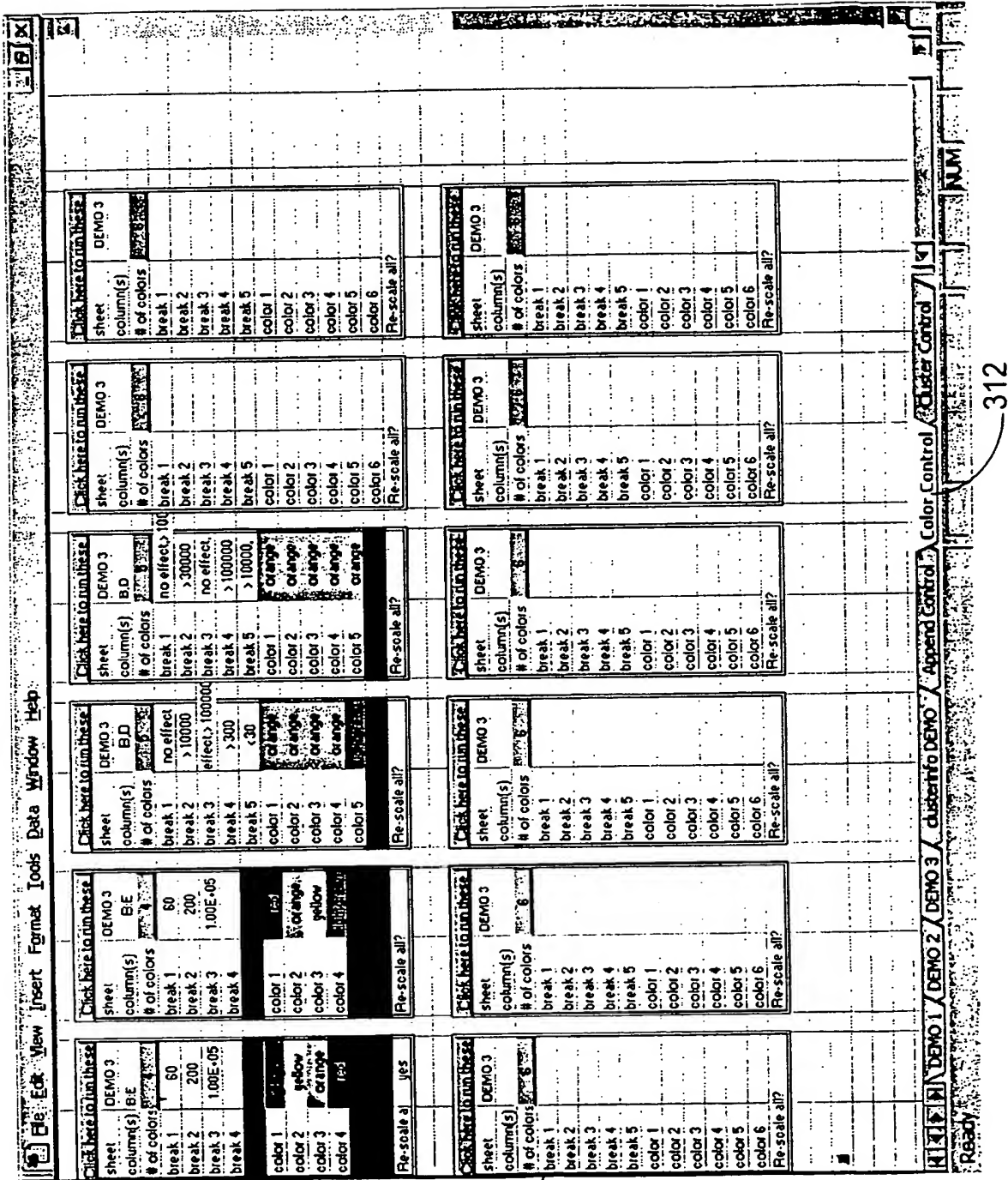


Fig. 4A

Fig. 4B



132 Click here to run these

sheet	DEMO 1
column(s)	E
# of colors	6
break 1	1
break 2	5
break 3	10
break 4	

140

134

136

138

color 1	light green
color 2	yellow
color 3	orange
color 4	red

142

130

Re-scale all:	no
---------------	----

Fig. 5A

130-1

Click here to run these	
sheet	DEMO 1
column(s)	C,D
# of colors	4
break 1	0.1
break 2	1
break 3	5
break 4	
color 1	light green
color 2	yellow
color 3	orange
color 4	red
Re-scale all	no

Fig. 5B

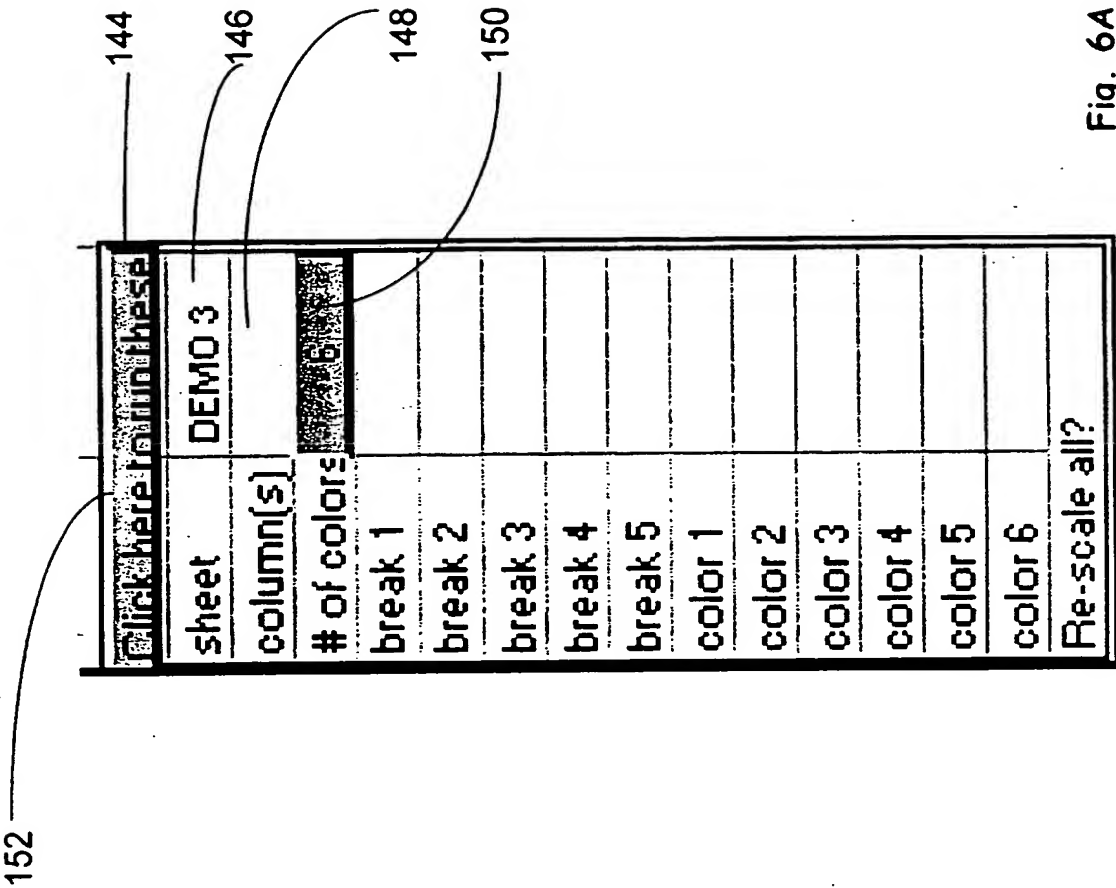


Fig. 6A

Fig. 6B

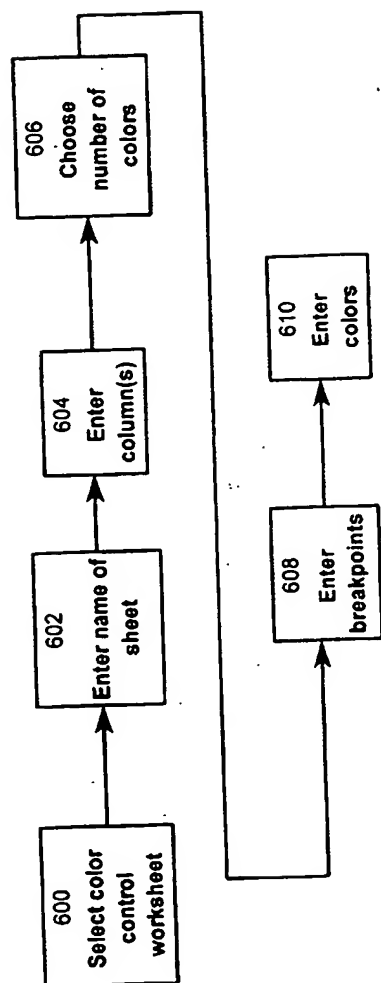
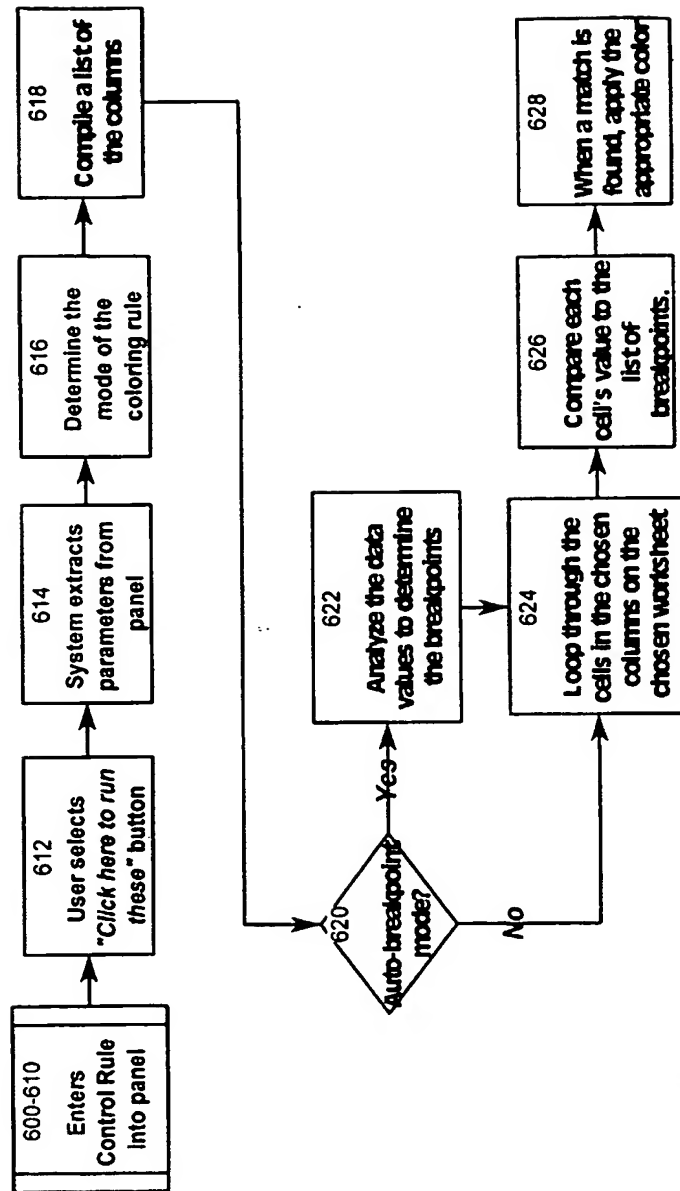


Fig. 6C



File Edit View Insert Format Tools Data Window Help														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M						
Cmpd01	N		23	30	41	3	22	5						
Cmpd02	N		42	55	83	57	28	15						
Cmpd03	G		23	11	70	23	24	29						
Cmpd04	N			30	89	60	21	22						
Cmpd05	N			92	71	41	13	3						
Cmpd06	D	8.85	6.5	3.7	100	79	49	43						
Cmpd07	D	1.11		7.8	65	28	28	38						
Cmpd08	D			2.6	68	41	22	15						
Cmpd09	D	0.119			61	42	24	5						
Cmpd10	N	0.233			50	77	63	25						
Cmpd11	N	1.41			47	25	24	3						
Cmpd12	H	1.3	0.24		81	59	4	37						
Cmpd13	H	1.18	0.194	30	39	23	4	12						
Cmpd14	H	0.26	0.41		99	46	46	36						
Cmpd15	H	0.369	0.148		101	82	38	18						
Cmpd16	I		0.87	30	81	64	47	24						
Cmpd17	K		0.223	30	79	54	22	32						
Cmpd18	I	5.27			71	71	23	12						
Cmpd19	I	0.104			101	109	108	100						
Cmpd20	F		0.317		87	70	31	13						
Cmpd21	K		2.21		94	77	36	12						
Cmpd22	B		0.15		96	61	36	12						
Cmpd23	B				110	91	69	39						
Cmpd24	B		0.27		105	104	75	52						
Cmpd25	B				93	71	41	22						
Cmpd26	B	0.663			97	79	43	23						
Cmpd27	B	0.111			95	93	52	25						
Cmpd28	E		0.13		68	62	12	11						
Cmpd29	J		0.46		41	48	9	17						
Cmpd30	J			5.2	5	5	21	8						
Cmpd31	N				62	30	15	33						
Cmpd32	J				112	29	75	14						
Cmpd33	N				105	62	62	21						
Cmpd34	J	2.27	24	30	105	62	62	21						
Cmpd35	J	0.83			44	28	18	3						

Fig. 7A

File Edit View Insert Format Tools Data Window Help													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-5M	HTS SPA Dose-Resp % Inhib @ 1x10-5M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M					
32	Cmpd32	J			62	30	15	33					
33	Cmpd33	N			112	29	79	14					
34	Cmpd34	J	0.23	30	105	62	62	21					
35	Cmpd35	J	0.63		44	28	18	3					
36	Cmpd36	J			92	63	42	14					
37	Cmpd37	N	0.23		58	20	38	30					
38	Cmpd38	J			90	60	51	2					
39	Cmpd39	L			87	73	29	15					
40	Cmpd39	F	0.56	0.41	85	56	40	11					
41	Cmpd40	O	0.368		66	46	28	13					
42	Cmpd41	G		0.36	102	87	92	87					
43	Cmpd42	F	0.28	0.9	36	25	18	12					
44	Cmpd43	H	0.13		76	60	40	25					
45	Cmpd44	O			111	110	104	82					
46	Cmpd45	F	0.6		25	104	16	21					
47	Cmpd46	M			109	104	97	80					
48	Cmpd47	F			100	102	83	75					
49	Cmpd48	F	0.27		71	43	31	15					
50	Cmpd49	M			112	111	112	75					
51	Cmpd50	F			78	70	44	27					
52	Cmpd51	O			69	24	23	14					
53	Cmpd52	H	0.33	4.9	108	102	63	27					
54	Cmpd53	A	0.18	1.6	91	78	76	33					
55	Cmpd54	A	0.33		61	64	60	26					
56	Cmpd55	O	0.12		93	80	55	37					
57	Cmpd56	C	0.39		101	72	42	29					
58	Cmpd57	A			92	69	55	49					
59	Cmpd58	C	0.22										
60													
61													
62													
63													
64													
65													
66													
67													
68													
69													
70													

Fig. 7B

File Edit View Insert Format Tools Data Window Help													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M					
Cmpd01	N		23	30		3	32	5					
Cmpd02	N		12	6.5	41	57	28	15					
Cmpd03	G		2.1	11	93			29					
Cmpd04	N			20	70	25	24	23					
Cmpd05	N		1.3	3.2	89	60	21	22					
Cmpd06	D	888	6.5	3.7	71		13	3					
Cmpd07	D	2.1	3.0	7.8		79	49	43					
Cmpd08	D		0.8	2.6	65	28	28	38					
Cmpd09	D	0.119			68	41	22	15					
Cmpd10	N	0.233			81		24	5					
Cmpd11	N	1.31			60	77	63	25					
Cmpd12	H	1.3	0.24		47	25	24	3					
Cmpd13	H	1.11	0.194	30	81	53	40	37					
Cmpd14	H	0.26	0.41		39	23	4	12					
Cmpd15	H	0.369	0.149			48	46	38					
Cmpd16	I		0.87	30		82	39	18					
Cmpd17	K		0.223	30	81	64	47	24					
Cmpd18	I	5.27			79	64	22	32					
Cmpd19	I	0.134			71	71	23	12					
Cmpd20	F		0.317			181							
Cmpd21	K		2.21		87	70	31	13					
Cmpd22	B		0.16			77	36	12					
Cmpd23	B					61	36	12					
Cmpd24	B		0.27			11	69	39					
Cmpd25	B						75	52					
Cmpd26	B	0.655				71	41	22					
Cmpd27	B	0.111				79	43	23					
Cmpd28	E		0.13				62	25					
Cmpd29	J		0.46		68	62	12	11					
Cmpd30	J				11	40	9	17					
Cmpd31	N		4.5		5	6	21	8					
Cmpd32	J				62	30	15	33					
Cmpd33	N					28	75	14					

Fig. 8A

Fig. 8B

File Edit View Insert Format Tools Data Window Help													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-5M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M					
Cmpd34	J	0.23	0.24	0.30	82	62	62	21					
Cmpd35	J	0.63			44	28	18	3					
Cmpd36	J		0.23		63	63	42	14					
Cmpd37	N	1.23			30	30	61	30					
Cmpd38	L	0.00			90	65	61	2					
Cmpd39	F	0.66	0.41		87	73	29	15					
Cmpd40	G	0.368	0.41		65	46	40	11					
Cmpd41	G	0.00	0.35		66	46	28	13					
Cmpd42	F	0.38	0.3			87		87					
Cmpd43	H	0.46	0.13		78	60	40	12					
Cmpd44	G	0.00				60	40	25					
Cmpd45	F	0.5				21	16	21					
Cmpd46	M	1.14			25			80					
Cmpd47	F	0.00					83	76					
Cmpd48	F	0.27			71	43	31	15					
Cmpd49	M	0.00						75					
Cmpd50	F	0.00			78	70	44	27					
Cmpd51	G			4.9	69	24	23	14					
Cmpd52	H	0.33	0.117				63	27					
Cmpd53	A	0.00	0.16				76	33					
Cmpd54	A	0.33				78	77	26					
Cmpd55	G	0.12			61	64	60	37					
Cmpd56	C	0.39				80	65	31					
Cmpd57	A	0.00				72	42	29					
Cmpd58	C	0.22				69	55	48					
Cmpd59													
Cmpd60													
Cmpd61													
Cmpd62													
Cmpd63													
Cmpd64													
Cmpd65													
Cmpd66													
Cmpd67													
Cmpd68													
Cmpd69													
Cmpd70													
Cmpd71													
Cmpd72													
Cmpd73													
Cmpd74													
Cmpd75													
Cmpd76													
Cmpd77													
Cmpd78													
Cmpd79													
Cmpd80													
Cmpd81													
Cmpd82													
Cmpd83													
Cmpd84													
Cmpd85													
Cmpd86													
Cmpd87													
Cmpd88													
Cmpd89													
Cmpd90													
Cmpd91													
Cmpd92													
Cmpd93													
Cmpd94													
Cmpd95													
Cmpd96													
Cmpd97													
Cmpd98													
Cmpd99													
Cmpd100													

[illegible]

Fig. 9A

File Edit View Insert Format Tools Data Window Help														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
45	F4 D	F4 D	22											
46	F5 A	F5 A	1											
47	F5 B	F5 B	21											
48	F5 C	F5 C	22											
49	F5 D	F5 D	22											
50	F6 A	F6 A	49											
51	F6 B	F6 B	35											
52	F6 C	F6 C	26											
53	F6 D	F6 D	14											
54	F7 A	F7 A	4											
55	F7 B	F7 B	19											
56	F7 C	F7 C	9											
57	F7 D	F7 D	15											
58														
59														
60														
61														
62														
63														
64														
65														
66														
67														
68														
69														
70														
71														
72														
73														
74														
75														
76														
77														
78														
79														
80														
81														
82														
83														
84														
85														
86														
87														

Fig. 9B

Microsoft Excel - demo~PANDORA~v23.1.xls [Read-Only]													
	A	B	C	D	E	F	G	H	I	J	K	L	M
	mixture	acid	amine	activity									
			group	@1µM									
223	E1A	E1	A	37									
224	E1B	E1	B	38									
225	E1C	E1	C	40									
226	E1D	E1	D	37									
227	E2A	E2	A	7									
228	E2B	E2	B	38									
229	E2C	E2	C	40									
230	E2D	E2	D	37									
231	E3A	E3	A	37									
232	E3B	E3	B	38									
233	E3C	E3	C	26									
234	E3D	E3	D	25									
235	E4A	E4	A	17									
236	E4B	E4	B	27									
237	E4C	E4	C	12									
238	E4D	E4	D	3									
239	E5A	E5	A	5									
240	E5B	E5	B	28									
241	E5C	E5	C	10									
242	E5D	E5	D	10									
243	E6A	E6	A	18									
244	E6B	E6	B	10									
245	E6C	E6	C	17									
246	E6D	E6	D	73									
247	E7A	E7	A	28									
248	E7B	E7	B	42									
249	E7C	E7	C	32									
250	E7D	E7	D	42									
251	F1A	F1	A	34									
252	F1B	F1	B	41									
253	F1C	F1	C	32									
254	F1D	F1	D	31									
255	F2A	F2	A	31									
256	F2B	F2	B	37									
257	F2C	F2	C	31									
258	F2D	F2	D	38									
259	F3A	F3	A	38									
260	F3B	F3	B	33									
261	F3C	F3	C	42									
262	F3D	F3	D	41									
263	F4A	F4	A	41									

Fig. 10A

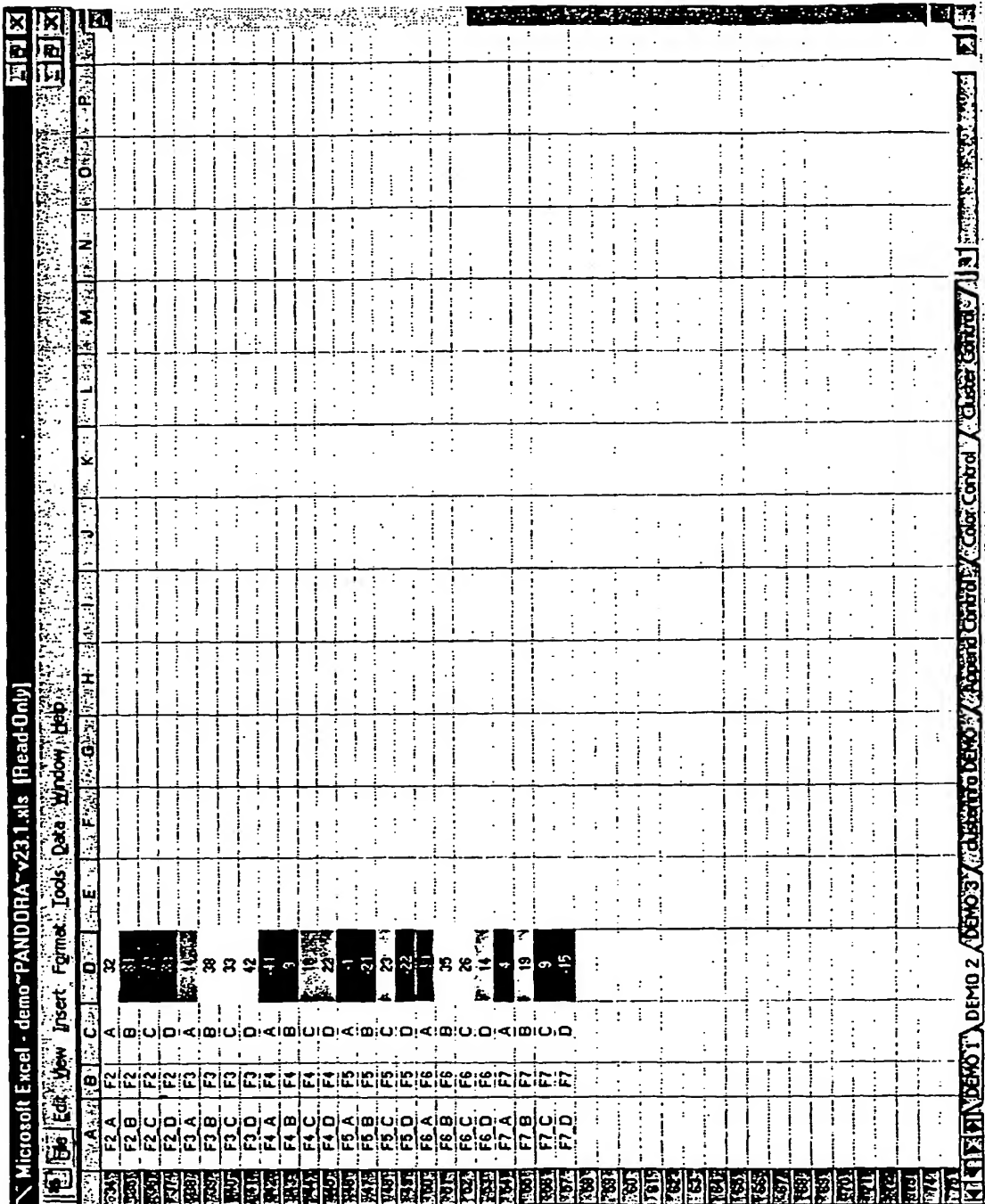


Fig. 10B

File Edit View Insert Format Tools Data Window Help

Name:	Cmd
Sheet #	DEMO 1
Cluster Col	A

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 2

Cluster Col C

Color	Score
red	0

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 3

Cluster Col B

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

File Edit View Insert Format Tools Data Window Help

Name:	Cmd
Sheet #	DEMO 1
Cluster Col	A

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 2

Cluster Col C

Color	Score
red	0

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 3

Cluster Col B

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

File Edit View Insert Format Tools Data Window Help

Name:	Cmd
Sheet #	DEMO 1
Cluster Col	A

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 2

Cluster Col C

Color	Score
red	0

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 3

Cluster Col B

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

File Edit View Insert Format Tools Data Window Help

Name:	Cmd
Sheet #	DEMO 1
Cluster Col	A

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 2

Cluster Col C

Color	Score
red	0

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 3

Cluster Col B

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

File Edit View Insert Format Tools Data Window Help

Name:	Cmd
Sheet #	DEMO 1
Cluster Col	A

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 2

Cluster Col C

Color	Score
red	0

Column #	Ref. Weight
C,D	1
E	1

Name: amine

Sheet # DEMO 3

Cluster Col B

Color	Score
red	1
yellow	1

Column #	Ref. Weight
C,D	1
E	1

Fig. 11A

314

1100

File Edit View Insert Format Tools Data Window Help

Insert Data Table

Columns F:1

Name: acids

Sheet # DEMO 2

Cluster Col B

Color red

Score 0

Rel. Weight 1

Insert Data Table

Columns F:1

Name: amines

Sheet # DEMO 2

Cluster Col C

Color yellow

Score 2

Rel. Weight 1

Insert Data Table

Columns F:1

Name: DEMO 3

Cluster Col

Color

Score

Rel. Weight

Cluster Control

Cluster Control A

Fig. 11B

Name: Grnd	
Sheet #	DEMO 1
Cluster Col	A
Color	
red	Score 1
orange	1
yellow	1
Columns	
C,D	Rel. Weight 1
E	1
Columns	
F,I	

Fig. 11D

Microsoft Excel - demo~PANDORA~v23.1.xls [Read-Only]														
File Edit View Insert Format Tools Data Window Help														
B1														
mixture	acid	amine group	activity @ 1µM											
E1A	E1	A												
E1B	E1	B												
E1C	E1	C	40											
E1D	E1	D												
E2A	E2	A	7											
E2B	E2	B												
E2C	E2	C	40											
E2D	E2	D												
E3A	E3	A	17											
E3B	E3	B	22											
E3C	E3	C	20											
E3D	E3	D	14											
E4A	E4	A	22											
E4B	E4	B	12											
E4C	E4	C	33											
E4D	E4	D	12											
E5A	E5	A	3											
E5B	E5	B	5											
E5C	E5	C	28											
E5D	E5	D	10											
E6A	E6	A	5											

Fig. 12

File Edit View Insert Format Tools Data Window Help													
COMP ID	# of replicates	S#	MPTS Test1	IC50 (nM) Test1	DATE Test1	MPTS Test2	IC50 (nM) Test2	DATE Test2	SELECT MITY	K	L	M	N
(-)-pentazocine	4			no effect > 10000	7/20/97	4	8384	4/10/97					
(-)-pentazocine	4			no effect > 10000	7/20/97	4	> 10000 (blank)	4/10/97					
(-)-pentazocine	4			no effect > 10000	7/20/97	4	3133	35530	0.00				
0277	3			165.3333333	3/6/97	3	(blank) > 10000	35498	0.00				
0278	3				4/3/97	3	3575	35498	38.21				
0629				no effect	10/8/97								
0651				188.6666667	3/6/97	3	4100	3/9/97	15.59				
0657				no effect	10/6/97								
0793					4/17/97	4	26885	35539.85714	679.87				
0958					8/13/97	5	1283	8/18/97	32.08				
1068					6/18/97	4	4000	6/23/97	8.99				
10817-018-C				no effect	7/22/98								
1210					6/19/97	6	75000	6/24/97	2080.23				
1329					8/13/97	5	23181	8/19/97	1158.05				
1336					8/6/97	5	32903	8/7/97	310.41				
1337				108	8/14/97	5	43105	8/20/97					
1338				> 300	8/6/97	5	> 100000	8/7/97					
1339				no effect	8/14/97	5	85442	8/19/97	251.61				
1341				no effect	8/14/97	5	> 100000	8/19/97					
1342				141	8/6/97	5	34184	8/7/97	242.44				
1343				149	7/23/97	5	62235	7/31/97	418.09				
1344				145	7/25/97	5	5080	7/30/97	35.03				
1362				> 300	7/1/97	4	53000	7/2/97	235.56				
1369				no effect	7/1/97	4	no effect	7/2/97					
1420				111	8/13/97	5	4657	8/18/97	41.95				
1431				no effect	7/25/97	5	16987	7/30/97					
1439				no effect	8/14/97	5	no effect	8/20/97					
1444				no effect	8/14/97	5	> 100000	8/20/97					
1445				no effect	10/8/97								
1446				no effect	8/14/97	5	no effect	8/20/97					
1447				no effect	10/8/97								
1453				> 300	7/25/97	5	15484	7/30/97	469.21				
1462				no effect	9/25/97								
1464				no effect	10/3/97								
1465				no effect	10/8/97								
1466				no effect	8/15/97	5	> 100000	8/21/97					
1467				no effect	9/18/97								
1468				no effect	9/25/97								
Ready													

Fig. 13A

[illegible]

Fig. 13B

306

318

A	B	C	D	E	F	G	H	I	J	K	L	M	N
CMPD ID	# of replicates	S#	#PTS Test1	ICS0 (nM) Test1	DATE Test1	#PTS Test2	ICS0 (nM) Test2	DATE Test2	SELECT WITY				
(-)pentazocine				no effect, 10000	7/20/97	4	8384	4/10/97					
(-)pentazocine				no effect, 10000	7/20/97	4	> 10000 (blank)	4/10/97					
(-)pentazocine				158.3333333	7/20/97	4	3133	35530	0.00				
0277					3/6/97	3	(blank), 10000	35498	0.00				
0278					4/3/97	3	3576	35498	38.21				
0629				no effect	10/9/97								
0661				158.6565657	3/6/97	3	4100	3/6/97	15.59				
0697				no effect	10/6/97								
0793					4/17/97	4	285714286	35539.85714	675.87				
0958					8/13/97	5	1293	8/18/97	32.09				
1058					6/18/97	4	4000	6/23/97	8.99				
10877-018-C				no effect	11/22/98								
1210													
1329													
1336													
1337													
1338													
1339													
1341													
1342													
1343													
1344													
1362					7/11/97	4	53000	7/28/97	235.56				
1369				> 300	7/11/97	4	no effect	7/28/97					
1420				111	8/13/97	5	4657	8/18/97	41.95				
1431				no effect	7/25/97	5	16367	7/30/97					
1439				no effect	8/14/97	5	no effect	8/20/97					
1444				no effect	8/14/97	5	> 100000	8/20/97					
1445				no effect	10/6/97								
1446				no effect	8/14/97	5	no effect	8/20/97					
1447				no effect	10/6/97								
1453				> 300	7/25/97	5	15484	7/30/97	463.21				
1463				no effect	9/25/97								
1464				no effect	10/3/97								
1465				no effect	10/8/97								
1466				no effect	8/15/97	5	> 100000	8/21/97					
1467				no effect	9/16/97								
1468				no effect	8/15/97								

ENTER SCALING FACTOR FOR DISPLAY

Enter a vertical scale factor, relative to current size.

OK Cancel

14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50

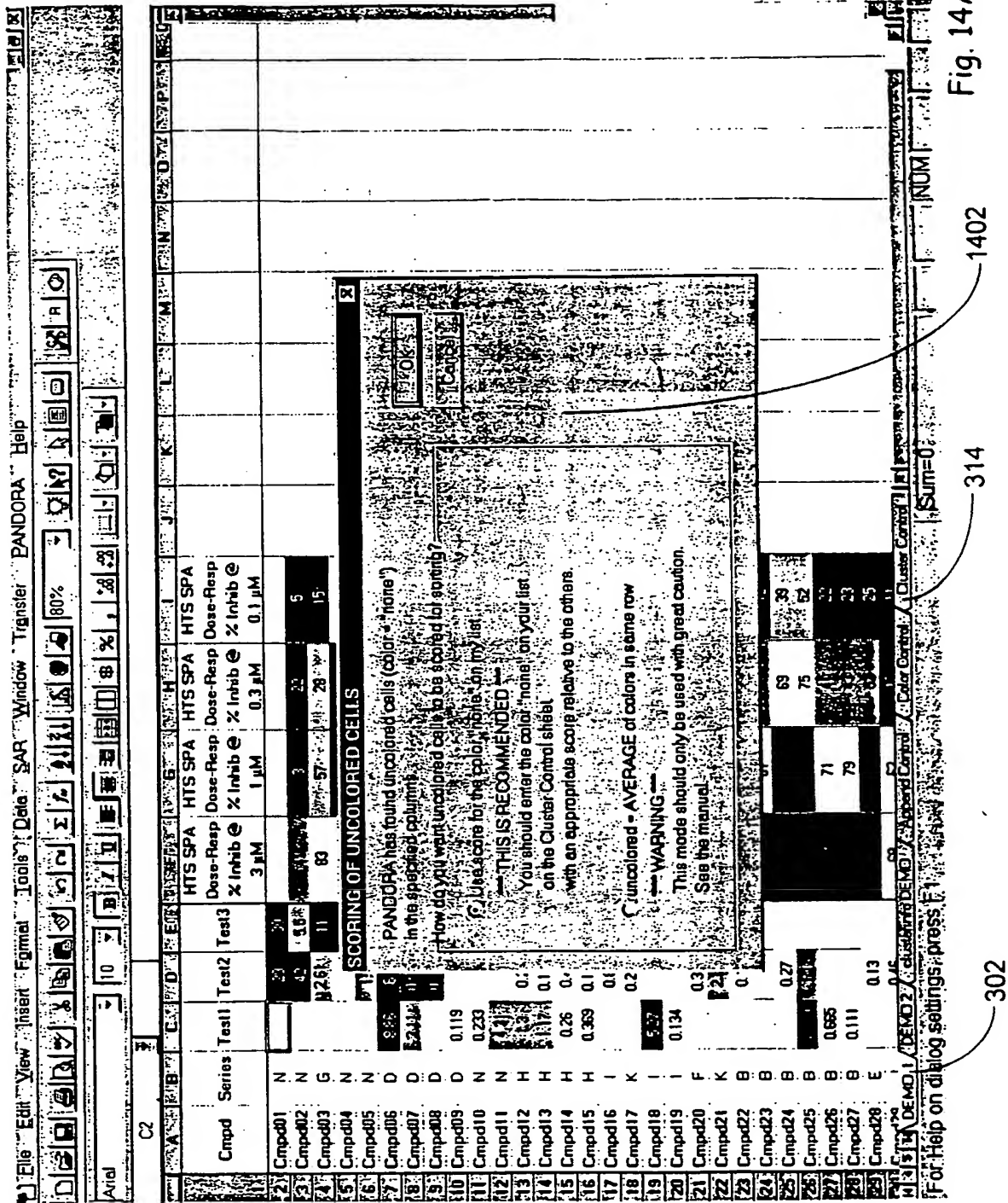
DEMO 1 / DEMO 2 / DEMO 3 / Append Control / Cluster Control / NUM

Fig. 13C

306

[illegible]

Fig. 13D



File Edit View Insert Format Tools Data Window Help													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M					
2	Cmpd38	L			80	60	51	2					
3	Cmpd44	G			78	60	40	25					
4	Cmpd49	M			71	63	31	15					
5	Cmpd53	A		1.6			63	27					
6	Cmpd09	D		2.6	65	54	79	28					
7	Cmpd41	G		0.35	68	66	25	13					
8	Cmpd52	H		0.33	69	24	23	13					
9	Cmpd07	D		2.4		79	48	43					
10	Cmpd14	H		0.26	0.41	28	4	12					
11	Cmpd15	H		0.363	0.148	23	4	12					
12	Cmpd26	B		0.57									
13	Cmpd39	F		0.56	0.41								
14	Cmpd42	F		0.39	0.9								
15	Cmpd47	F											
16	Cmpd50	F											
17	Cmpd51	G											
18	Cmpd57	A											
19	Cmpd09	D		0.119									
20	Cmpd10	N		0.233									
21	Cmpd12	H			0.24								
22	Cmpd19	I		0.134									
23	Cmpd20	F			0.317								
24	Cmpd22	B			0.16								
25	Cmpd24	B			0.27								
26	Cmpd26	B		0.665			69	22					
27	Cmpd27	B		0.111		71	11	23					
28	Cmpd28	E				79	62	25					
29	Cmpd29	J		0.13			12	11					
30	Cmpd35	J		0.63		68	18	3					
31	Cmpd36	J				44	63	14					
32	Cmpd40	G				65	40	11					
33	Cmpd43	H		0.368		38	12	12					
34	Cmpd45	F		0.13		25		82					
35				0.5									
Microsoft Excel													
DONE!													
Summary: 58 rows in 58 columns													
Elapsed time: 5 seconds													
36 columns loaded in seconds													
Two sheets have been added for convenience in your workbook:													
"DEMO1 SCORES by Cmpd"													
"DEMO1 SORTED by Cmpd un-zse"													
OK													
DEMO1 X DEMO1 SCORES by Cmpd DEMO1 SORTED by Cmpd un-zse DEMO2 DEMO3 X DEMO3 X													

Fig. 14B

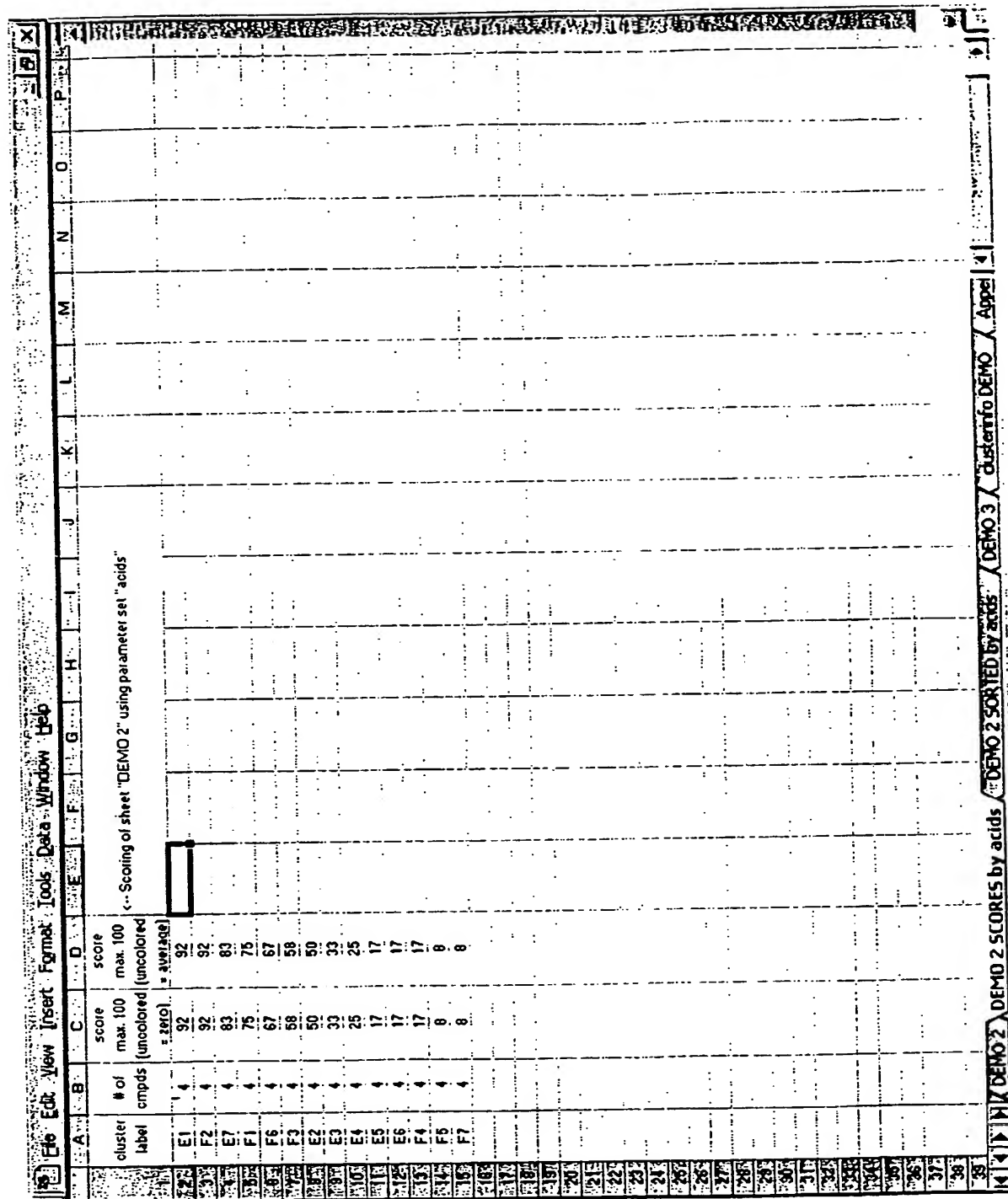
Fig. 14C

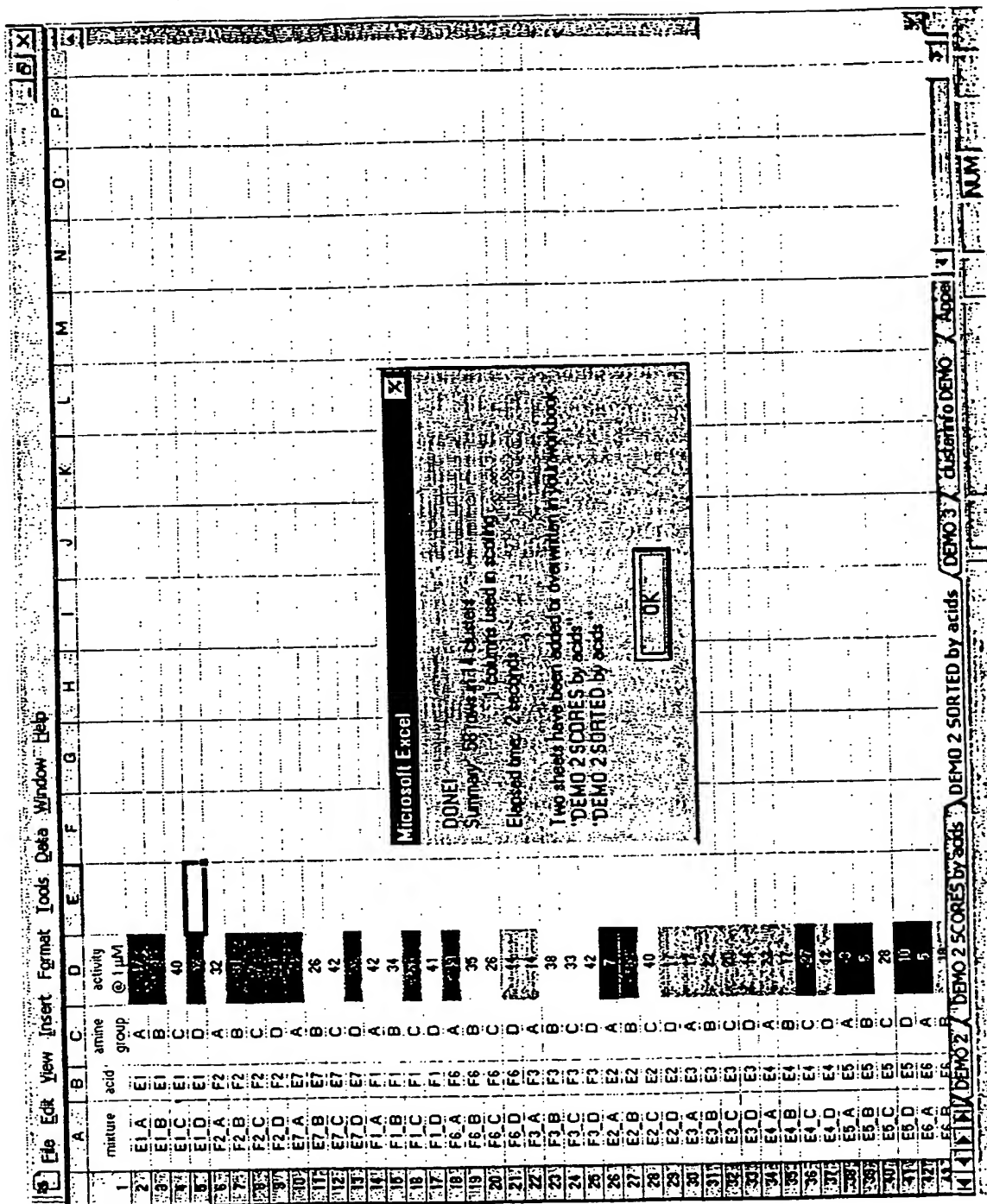
File Edit View Insert Format Tools Data Window Help																											
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P												
cluster label	# of cmpds	score max. 100 (uncolored = zero)	max. 100 (uncolored = zero)	← Scaling of sheet "DEMO 1" using parameter set "Cmpd"																							
1																											
20	Cmpd38	1	67	100																							
30	Cmpd44	1	67	100																							
40	Cmpd49	1	67	100																							
50	Cmpd53	1	67	50																							
60	Cmpd08	1	50	75																							
70	Cmpd41	1	50	50																							
80	Cmpd52	1	50	50																							
90	Cmpd07	1	33	25																							
100	Cmpd14	1	33	50																							
110	Cmpd15	1	33	50																							
120	Cmpd25	1	33	50																							
130	Cmpd39	1	33	50																							
140	Cmpd42	1	33	50																							
150	Cmpd47	1	33	100																							
160	Cmpd50	1	33	50																							
170	Cmpd51	1	33	50																							
180	Cmpd57	1	33	100																							
190	Cmpd09	1	17	50																							
200	Cmpd10	1	17	50																							
210	Cmpd12	1	17	50																							
220	Cmpd19	1	17	50																							
230	Cmpd20	1	17	50																							
240	Cmpd22	1	17	50																							
250	Cmpd24	1	17	50																							
260	Cmpd26	1	17	50																							
270	Cmpd27	1	17	50																							
280	Cmpd28	1	17	50																							
290	Cmpd29	1	17	50																							
300	Cmpd35	1	17	50																							
310	Cmpd36	1	17	50																							
320	Cmpd40	1	17	50																							
330	Cmpd43	1	17	50																							
340	Cmpd45	1	17	50																							
350	Cmpd48	1	17	50																							
360	Cmpd54	1	17	50																							
370	Cmpd56	1	17	50																							
DEMO 1 SCORES by Cmpd / DEMO 1 SORTED by Cmpd unsize / DEMO 2 / DEMO 3 / cluster																											
Ready																											

Ready

File Edit View Insert Format Tools Data Window Help																
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
30	Cmpd35	17	55													
31	Cmpd36	17	55													
32	Cmpd40	17	55													
33	Cmpd43	17	55													
34	Cmpd45	17	55													
35	Cmpd48	17	55													
36	Cmpd54	17	55													
37	Cmpd56	17	55													
38	Cmpd58	17	55													
39	Cmpd05	0	0													
40	Cmpd11	0	0													
41	Cmpd13	0	0													
42	Cmpd16	0	0													
43	Cmpd17	0	0													
44	Cmpd21	0	0													
45	Cmpd23	0	0													
46	Cmpd31	0	0													
47	Cmpd32	0	0													
48	Cmpd33	0	0													
49	Cmpd46	0	0													
50	Cmpd02	-33	-33													
51	Cmpd03	-33	-33													
52	Cmpd04	-33	-33													
53	Cmpd06	-33	-33													
54	Cmpd18	-33	-33													
55	Cmpd30	-33	-33													
56	Cmpd37	-33	-33													
57	Cmpd55	-33	-33													
58	Cmpd01	-67	-67													
59	Cmpd34	-67	-67													
60																
61																
62																
63																
64																
65																
66																
67																
68																
69																
70																
71																

Fig. 14D





	A	B	C	D	E
	Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M
1					
2	Cmpd01	90	22	16	19
3	Cmpd02	41	3	22	5
4	Cmpd03	83	57	28	15
5	Cmpd04	70	25	24	29
6	Cmpd05	89	60	21	22
7	Cmpd06	71	41	13	3
8	Cmpd07	100	70	48	43
9	Cmpd08	65	28	28	38
10	Cmpd09	88	41	22	15
11	Cmpd10	61	42	24	5
12	Cmpd11	50	77	63	25
13	Cmpd12	47	25	24	3
14	Cmpd13	81	59	40	37
15	Cmpd14	39	23	4	12
16	Cmpd15	99	46	46	36
17	Cmpd16	100	90	38	18
18	Cmpd17	91	64	47	24
19	Cmpd18	79	54	22	32
20	Cmpd19	71	71	23	12
21	Cmpd20	100	100	100	100

Figure 16A

Click here to run these	
sheet	DEMO 1
column(s)	B:E
# of colors	3
break 1	33
break 2	67
break 3	
color 1	red
color 2	yellow
color 3	light green
Re-scale all?	

Figure 16B

A	B	C	D	E	F	G	H
Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁶ M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁶ M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁷ M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁷ M 1.00e-07	D-R -iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd
1							
2	100	100	100	100	75	100	100
3	100	59	48	43	83	77	79
4	81	59	40	37	83	70	75
5	90	46	46	36	83	70	75
6	100		38	18	92	50	70
7	81	64	47	24	92	43	67
8	81	57	28	15	92	23	57
9	80	60	21	22	92	23	57
10	71	41	13	3	92	23	57
11	80	41	22	15	92	23	57
12	50		63	25	67	47	57
13	71	54	22	32	92	23	57
14	71		23	12	83	30	56
15	61	42	24	5	83	20	51
16	90	22	16	19	83	10	46
17	71	25	24	29	83	10	46
18	41	3	22	5	83	7	45
19	65	28	28	38	58	33	45
20	47	25	24	3	83	7	45
21	39	23	4	12	83	7	45

Fig. 16C

	A	B	C	D	E	F	G	H
	Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10-6M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10-6M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10-7M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10-7M 1.00e-07	D-R iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composit score (0 to 100) by Cmpd
1								
22	marker_7.5	97	97	90	75	75	100	100
23	marker_7.0	97	91	75	50	83	87	86
24	marker_6.5	90	75	49	24	92	50	70
25	marker_6.0	75	50	23	9	92	23	57
26	marker_5.5	49	24	9	3	83	7	45
27	marker_5.0	23	9	3	1	75	0	38

Figure 16D

A	B	C	D	E	F	G	H	I	J
Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁶ M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁶ M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁷ M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁷ M 1.00e-07	D-R -iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd	Interp -log IC50 by Cmpd	est IC50 μM by Cmpd
1									
2	Cmpd20			100	75	100	100		<0.1
3	marker_7.5			100	75	100	100		
4	marker_7.0			50	83	87	86		
5	Cmpd07		48	43	83	77	79	6.78	0.17
6	Cmpd13	59	40	37	83	70	75	6.66	0.22
7	Cmpd15	46	46	36	83	70	75	6.66	0.22
8	Cmpd16		38	18	92	50	70	6.50	0.32
9	marker_6.5		49	24	92	50	70		
10	Cmpd17	64	47	24	92	43	67	6.38	0.41
11	Cmpd03	57	28	15	92	23	57	6.00	1
12	Cmpd05	60	21	22	92	23	57	6.00	1
13	Cmpd06	41	13	3	92	23	57	6.00	1
14	Cmpd09	41	22	15	92	23	57	6.00	1
15	Cmpd11		63	25	67	47	57	6.00	1
16	Cmpd18	54	22	32	92	23	57	6.00	1
17	marker_6.0	50	23	9	92	23	57		
18	Cmpd19		23	12	83	30	56	5.96	1.1
19	Cmpd10	42	24	5	83	20	51	5.75	1.8
20	Cmpd01	22	16	19	83	10	46	5.54	2.9
21	Cmpd04	25	24	29	83	10	46	5.54	2.9
22	Cmpd02	3	22	5	83	7	45		>3
23	Cmpd08	65	28	38	58	33	45		>3
24	Cmpd12	47	24	3	83	7	45		>3
25	Cmpd14	39	4	12	83	7	45		>3
26	marker_5.5	49	9	3	83	7	45		>3
27	marker_5.0	23	3	1	75	0	38		

Figure 16E

Fig. 16F

Table 2. The complete data set for 3 points and 3 colors, in systematic order.

compound	percent inhibition		data group number		data group color	
	highest conc	lowest conc	highest conc	lowest conc	highest conc	lowest conc
cmpd 01	2	29	1	1		
cmpd 02	15	10	1	2		
cmpd 03	31	26	1	3		
cmpd 04	21	46	1	1		
cmpd 05	30	53	2	2		
cmpd 06	17	37	2	3		
cmpd 07	26	90	1	1		
cmpd 08	10	90	3	2		
cmpd 09	32	72	3	3		
cmpd 10	34	17	1	1		
cmpd 11	51	8	2	2		
cmpd 12	56	3	2	3		
cmpd 13	33	39	2	1		
cmpd 14	53	52	2	2		
cmpd 15	51	52	2	3		
cmpd 16	65	82	2	1		
cmpd 17	43	71	2	2		
cmpd 18	65	99	2	3		
cmpd 19	67	11	3	1		
cmpd 20	87	5	3	2		
cmpd 21	77	8	3	3		
cmpd 22	78	36	3	1		
cmpd 23	85	40	3	2		
cmpd 24	83	57	3	3		
cmpd 25	73	88	3	1		
cmpd 26	69	85	3	2		
cmpd 27	79	68	3	3		

Fig. 166

Table 3. The complete data set for 3 points and 3 colors, sorted by decreasing dose-responsiveness

compound	highest conc	---->	lowest conc	step scoring	unscaled score points	scaled responsiveness 0-100
cmpd 22				+1+1	2	100
cmpd 10				+1+0	1	88
cmpd 13				0+1	1	88
cmpd 19				+1+0	1	88
cmpd 23				+1+0	1	88
cmpd 25				0+1	1	88
cmpd 26				0+1	1	88
cmpd 01				0+0	0	75
cmpd 14				0+0	0	75
cmpd 27				0+0	0	75
cmpd 04				-3+1	-2	50
cmpd 07				-3+1	-2	50
cmpd 08				-3+1	-2	50
cmpd 11				+1-3	-2	50
cmpd 12				+1-3	-2	50
cmpd 16				-3+1	-2	50
cmpd 17				-3+1	-2	50
cmpd 20				+1-3	-2	50
cmpd 21				+1-3	-2	50
cmpd 24				+1-3	-2	50
cmpd 02				0-3	-3	38
cmpd 03				0-3	-3	38
cmpd 05				-3+0	-3	38
cmpd 09				-3+0	-3	38
cmpd 15				0-3	-3	38
cmpd 18				-3+0	-3	38
cmpd 06				-3-3	-6	0

Fig. 16H

Table 4. The complete set of data for 3 points and 3 colors, sorted by decreasing overall activity.

compound	data group number		data group color		activity scoring	unscaled activity points	scaled activity 0-100
	highest conc	lowest conc	highest conc	lowest conc			
cmpd 27	3	3			1(3)+2(3)+3(3)	18	100
cmpd 18	2	3			1(2)+2(3)+3(3)	17	92
cmpd 24	3	3			1(3)+2(2)+3(3)	16	83
cmpd 09	1	3			1(1)+2(3)+3(3)	16	83
cmpd 26	3	2			1(3)+2(3)+3(2)	15	75
cmpd 15	2	3			1(2)+2(3)+3(3)	15	75
cmpd 17	2	2			1(2)+2(3)+3(2)	14	67
cmpd 21	3	1			1(3)+2(1)+3(3)	14	67
cmpd 06	1	2			1(1)+2(2)+3(3)	14	67
cmpd 23	3	2			1(3)+2(2)+3(2)	13	58
cmpd 08	1	3			1(1)+2(3)+3(2)	13	58
cmpd 12	2	1			1(2)+2(1)+3(3)	13	58
cmpd 25	3	1			1(3)+2(3)+3(1)	12	50
cmpd 14	2	2			1(2)+2(2)+3(2)	12	50
cmpd 03	1	1			1(1)+2(1)+3(3)	12	50
cmpd 16	2	3			1(2)+2(3)+3(1)	11	42
cmpd 20	3	1			1(3)+2(1)+3(2)	11	42
cmpd 05	1	2			1(1)+2(2)+3(2)	11	42
cmpd 22	3	2			1(3)+2(2)+3(1)	10	33
cmpd 07	1	3			1(1)+2(3)+3(1)	10	33
cmpd 11	2	1			1(2)+2(1)+3(2)	10	33
cmpd 13	2	2			1(2)+2(2)+3(1)	9	25
cmpd 02	1	1			1(1)+2(1)+3(2)	9	25
cmpd 19	3	1			1(3)+2(1)+3(1)	8	17
cmpd 04	1	2			1(1)+2(2)+3(1)	8	17
cmpd 10	2	1			1(2)+2(1)+3(1)	7	8
cmpd 01	1	1			1(1)+2(1)+3(1)	6	0

Fig. 16I

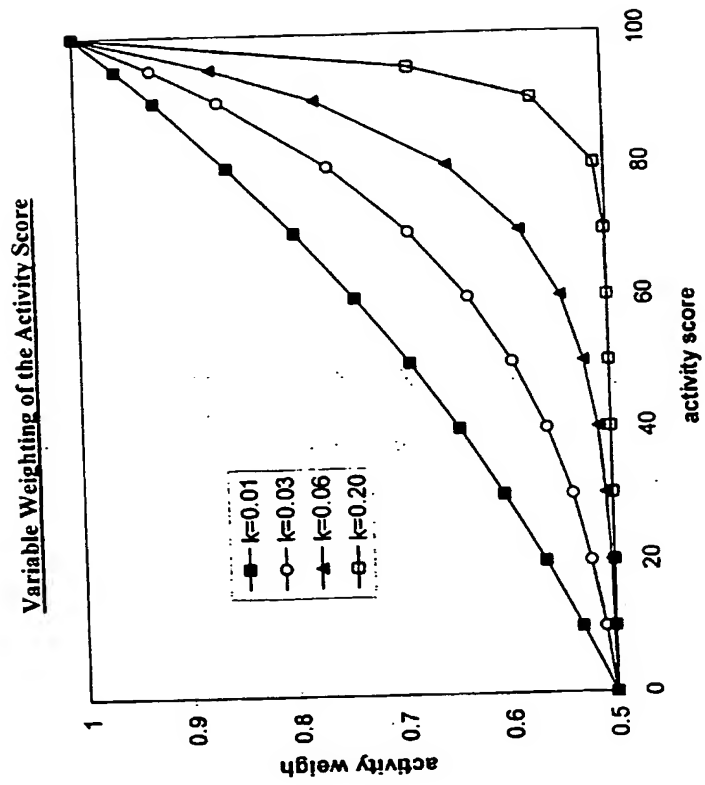


Fig. 16J

Table 5. The complete set of data for 3 points and 3 colors, sorted by decreasing composite score.

compound	highest conc	lowest conc	scaled responsive- ness 0-100	scaled activity 0-100	composite 0-100
cmpd 27			75	100	100
cmpd 18			38	92	82
cmpd 26			88	75	80
cmpd 24			50	83	72
cmpd 23			88	58	72
cmpd 09			38	83	69
cmpd 25			88	50	68
cmpd 22			100	33	66
cmpd 14			75	50	62
cmpd 15			38	75	61
cmpd 17			50	67	60
cmpd 21			50	67	60
cmpd 13			88	25	56
cmpd 08			50	58	54
cmpd 12			50	58	54
cmpd 19			88	17	52
cmpd 10			88	8	48
cmpd 16			50	42	46
cmpd 20			50	42	46
cmpd 03			38	50	44
cmpd 07			50	33	41
cmpd 11			50	33	41
cmpd 05			38	42	40
cmpd 06			0	67	38
cmpd 01			75	0	38
cmpd 04			50	17	33
cmpd 02			38	25	31

Fig. 16K

Quality of Ranking when Noise = 10 Inhibition Percentage Points

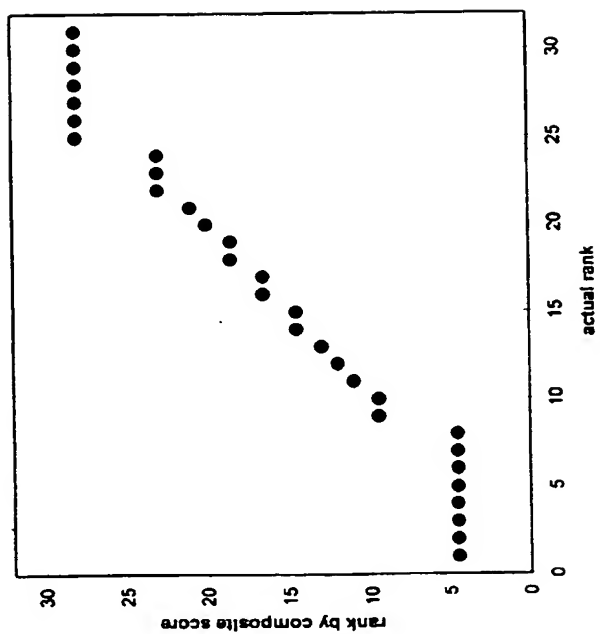


Fig. 16L

Quality of Ranking when Noise = 30 Inhibition Percentage Points

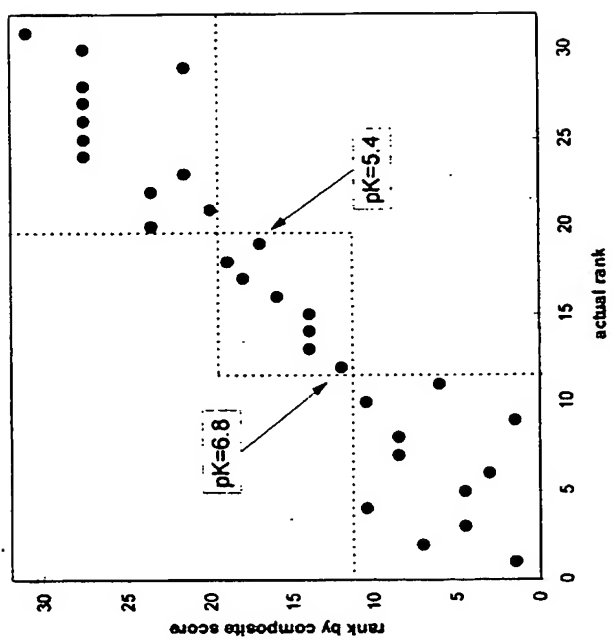


Fig. 16M

Table 6. Quantitative Estimation of Potencies by Calibration Marker Compounds

compound	% Inhib @ 3.00e-06	% Inhib @ 1.00e-06	% Inhib @ 3.00e-07	% Inhib @ 1.00e-07	D-R composite score	Interp -log IC50	est IC50 μM
M353875	100	102	83	75	79		<0.1
marker_7.5	99	97	90	76	79		
M221211	110	91	69	39	76	7.00	0.10
M371585	108	102	83	27	76	7.00	0.10
marker_7.0	97	91	75	50	76		
M345077	102	87	92	87	75	6.97	0.11
M371796	91	78	76	33	73	6.90	0.13
M143629	100	79	48	43	69	6.77	0.17
M371890	101	72	42	29	69	6.77	0.17
M371891	92	69	55	49	69	6.77	0.17
M309032	105	82	82	21	67	6.70	0.20
M198289	101	82	38	18	66	6.67	0.22
M224602	97	79	43	23	66	6.67	0.22
M318671	93	63	42	14	66	6.67	0.22
M273373	95	93	52	25	65	6.63	0.23
M371336	78	70	44	27	62	6.53	0.29
M371825	61	64	60	26	62	6.53	0.29
M181250	99	46	46	36	61	6.50	0.32
M338331	87	73	29	15	61	6.50	0.32
marker_6.5	90	76	49	24	61		
M143630	65	28	28	38	60	6.44	0.36

Fig. 16N

Quality of Estimation when Noise = 10 Inhibition Percentage Points

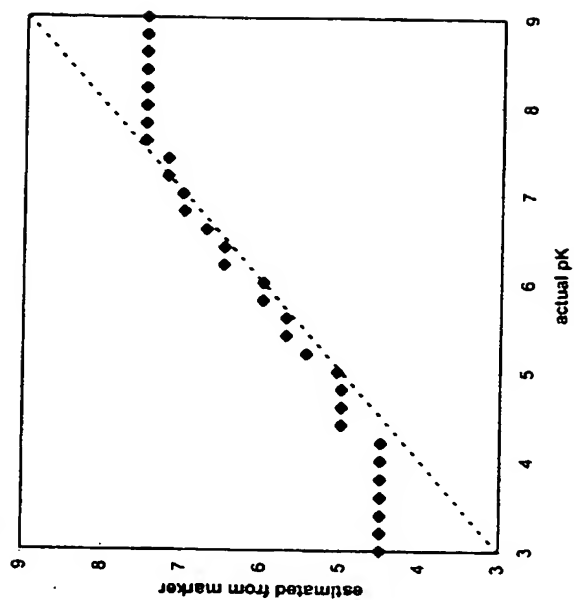


Fig. 16P

Quality of Estimation when Noise = 30 Inhibition Percentage Points

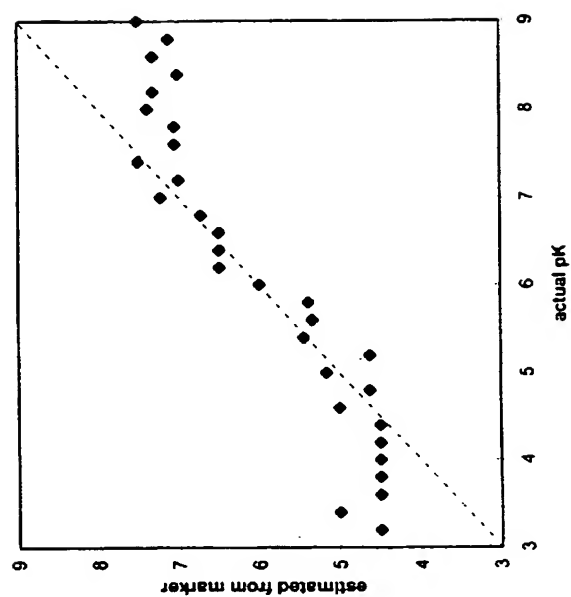
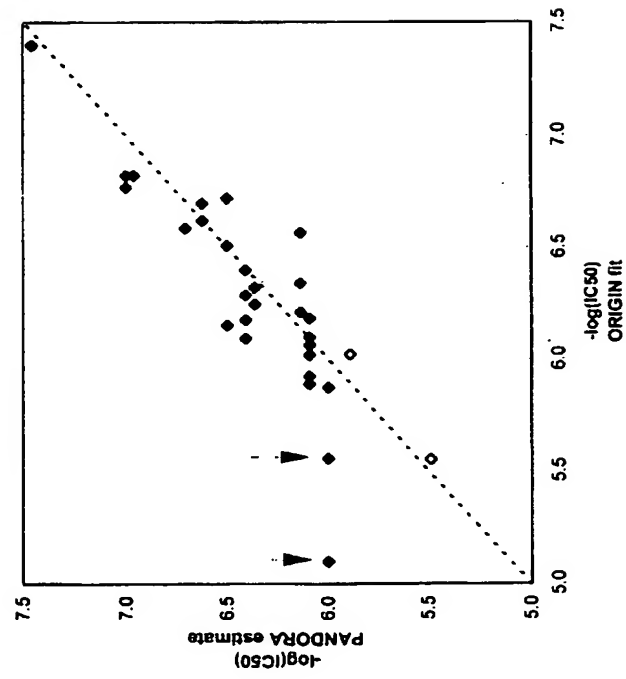


Fig. 16Q

Comparison of Curve Fitting to Marker Calibration
for T-cell Proliferation Data



A	B	C	D	E	F	G	H	I
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M
1								
2	Cmpd01	N	29	30	41	3	22	5
3	Cmpd02	N	42	55	83	57	28	15
4	Cmpd03	G	261	11	70	25	24	29
5	Cmpd04	N		30	89	60	21	22
6	Cmpd05	N	1.8	9.2	71	41	13	3
7	Cmpd06	D	6.5	3.7	100	79	48	43
8	Cmpd07	D	0.037	7.8	65	28	28	38
9	Cmpd08	D	0.089	N.A.	68	41	22	15
10	Cmpd09	D	0.119		61	42	24	5
11	Cmpd10	N	0.233		50	77	63	25
12	Cmpd11	N	431		47	25	24	3
13	Cmpd12	H	1.3	0.24	81	59	40	37
14	Cmpd13	H	1.17	0.194	39	23	4	12
15	Cmpd14	H	0.26	0.41	99	46	46	36
16	Cmpd15	H	0.369	0.148	101	82	38	18
17	Cmpd16			0.87	81	64	47	24
18	Cmpd17	K		0.223	79	54	22	32
19	Cmpd18		527		71	71	23	12
20	Cmpd19		0.134		101	109	108	100
21	Cmpd20			0.317	87	70	31	13
22	Cmpd21	K		2.21	94	77	36	12
23	Cmpd22	B		0.15	96	61	36	12
24	Cmpd23	B			110	91	69	39
25	Cmpd24	B	3.487	0.4				

Figure 17A

* THIS COLORING INDICATES A DATA COLUMN WITH MIXED DATA TYPES												
orig col	heading	# numeric	# text	# date	# blank	# total (longest col)	last occupied row num.	minimum (4 sig fig)	maximum (4 sig fig)	mean (4 sig fig)	standard dev (4 sig fig)	unique text strings and counts {24 different}
A	Cmpd	24				24	25					B(3) D(4) G(1) H(4) K(2) N(6)
B	Series		20			4	25					
C	Test1	12				12	25	0.119	8.86	2.385	2.726	
D	Test2	17				7	25	0.037	42	5.122	11.77	
E	Test3	10	2			12	25	0.4	30	15.76	12.59	NA (2)
F	HTS SPA Dose-Resp % Inhib @ 3x10-6M	23				1	25	39	110	77.57	20.43	
G	HTS SPA Dose-Resp % Inhib @ 1x10-6M	23				1	25	3	109	55.87	25.24	
H	HTS SPA Dose-Resp % Inhib @ 3x10-7M	23				1	25	4	108	35.52	21.85	
I	HTS SPA Dose-Resp % Inhib @ 1x10-7M	23				1	25	3	100	23.91	20.74	

Figure 17B

	A	B	C	D	E	F	G
	project name	most important factor scored by Mngr A	most important factor scored by Mngr B	most important factor scored by Mngr C	less important factor scored by Mngr A	less important factor scored by Mngr B	less important factor scored by Mngr C
1							
2	Proj 01	2		2	1	2	2
3	Proj 02	1	1	1	2	1	2
4	Proj 03	1	1	1	1	2	1
5	Proj 04	1	1	1	1	2	1
6	Proj 05	1	1	1	1	1	2
7	Proj 06	2		1	1	2	1
8	Proj 07	1	1	1	1	1	1
9	Proj 08	1	1	2	1	1	1
10	Proj 09	1	1	1	1	2	1
11	Proj 10	2	1	1	1	1	1
12	Proj 11	1	1	1	2	1	1
13	Proj 12	1	1	1	1	1	2
14	Proj 13	1	1	1	1	1	2
15	Proj 14	1	2	2	1	1	1
16	Proj 15	2		2	2	2	2
17	Proj 16	1	2	2	1	2	1
18	Proj 17	2	1		2	2	2
19	Proj 18	1			1		1
20	Proj 19	2	2	2	1	1	1
21	Proj 20	1	2	1	1	1	2

Figure 18A

Click here to run these	
sheet	Portfolio
columnn(s)	B:G
# of colors	3
break 1	1
break 2	2
break 3	3
color 1	red
color 2	yellow
color 3	green
Re-scale all?	

Figure 18B

Name: Factors	
Enlarge Cluster Starts	
Sheet #	Portfolio
Cluster Col	A
Shrink Cluster Starts	
Color	Score
red	1
yellow	2
Score and Sort Clusters	
Column(s)	Rel. Weight
B:D	3
E:G	1

Figure 18C

	A	B	C	D	E	F	G	H
	project name	most important factor scored by Mngr A	most important factor scored by Mngr B	most important factor scored by Mngr C	less important factor scored by Mngr A	less important factor scored by Mngr B	less important factor scored by Mngr C	score (0-100)
1								
2	Proj 18				1			94
3	Proj 05					1	2	92
4	Proj 04					2		86
5	Proj 14		2	2				83
6	Proj 20		2				2	83
7	Proj 13						2	81
8	Proj 09					2		75
9	Proj 12						2	75
10	Proj 15	2		2	2	2		75
11	Proj 01	2		2		2	2	72
12	Proj 08			2				69
13	Proj 06	2				2		67
14	Proj 17	2				2	2	67
15	Proj 19	2	2	2				64
16	Proj 07							61
17	Proj 03					2		58
18	Proj 10	2						58
19	Proj 16		2	2		2		58
20	Proj 02				2		2	39
21	Proj 11				2			36

Figure 18D

Fig. 19

Cmpd	Cluster	CD45	IRP	APOP	ACC	DON	cLogP	MW	HTS	Ratio	HTS	Ratio	Port
		HTS	HTS	HTS					Selectivity	A/Num	Ratio	Crit	
12142824	198						1.53	310.94	0.99	1.00			59.00
12142825	198						1.53	310.94	0.99	1.00			59.00
12142826	198						1.53	310.94	0.99	1.00			59.00
12142827	198						1.53	310.94	0.99	1.00			59.00
12142828	198						1.53	310.94	0.99	1.00			59.00
12142829	198						1.53	310.94	0.99	1.00			59.00
12142830	198						1.53	310.94	0.99	1.00			59.00
12142831	198						1.53	310.94	0.99	1.00			59.00
12142832	198						1.53	310.94	0.99	1.00			59.00
12142833	198						1.53	310.94	0.99	1.00			59.00
12142834	198						1.53	310.94	0.99	1.00			59.00
12142835	198						1.53	310.94	0.99	1.00			59.00
12142836	198						1.53	310.94	0.99	1.00			59.00
12142837	198						1.53	310.94	0.99	1.00			59.00
12142838	198						1.53	310.94	0.99	1.00			59.00
12142839	198						1.53	310.94	0.99	1.00			59.00
12142840	198						1.53	310.94	0.99	1.00			59.00
12142841	198						1.53	310.94	0.99	1.00			59.00
12142842	198						1.53	310.94	0.99	1.00			59.00
12142843	198						1.53	310.94	0.99	1.00			59.00
12142844	198						1.53	310.94	0.99	1.00			59.00
12142845	198						1.53	310.94	0.99	1.00			59.00
12142846	198						1.53	310.94	0.99	1.00			59.00
12142847	198						1.53	310.94	0.99	1.00			59.00
12142848	198						1.53	310.94	0.99	1.00			59.00
12142849	198						1.53	310.94	0.99	1.00			59.00
12142850	198						1.53	310.94	0.99	1.00			59.00
12142851	198						1.53	310.94	0.99	1.00			59.00
12142852	198						1.53	310.94	0.99	1.00			59.00
12142853	198						1.53	310.94	0.99	1.00			59.00
12142854	198						1.53	310.94	0.99	1.00			59.00
12142855	198						1.53	310.94	0.99	1.00			59.00
12142856	198						1.53	310.94	0.99	1.00			59.00
12142857	198						1.53	310.94	0.99	1.00			59.00
12142858	198						1.53	310.94	0.99	1.00			59.00
12142859	198						1.53	310.94	0.99	1.00			59.00
12142860	198						1.53	310.94	0.99	1.00			59.00
12142861	198						1.53	310.94	0.99	1.00			59.00
12142862	198						1.53	310.94	0.99	1.00			59.00
12142863	198						1.53	310.94	0.99	1.00			59.00
12142864	198						1.53	310.94	0.99	1.00			59.00
12142865	198						1.53	310.94	0.99	1.00			59.00
12142866	198						1.53	310.94	0.99	1.00			59.00
12142867	198						1.53	310.94	0.99	1.00			59.00
12142868	198						1.53	310.94	0.99	1.00			59.00
12142869	198						1.53	310.94	0.99	1.00			59.00
12142870	198						1.53	310.94	0.99	1.00			59.00
12142871	198						1.53	310.94	0.99	1.00			59.00
12142872	198						1.53	310.94	0.99	1.00			59.00
12142873	198						1.53	310.94	0.99	1.00			59.00
12142874	198						1.53	310.94	0.99	1.00			59.00
12142875	198						1.53	310.94	0.99	1.00			59.00
12142876	198						1.53	310.94	0.99	1.00			59.00
12142877	198						1.53	310.94	0.99	1.00			59.00
12142878	198						1.53	310.94	0.99	1.00			59.00
12142879	198						1.53	310.94	0.99	1.00			59.00
12142880	198						1.53	310.94	0.99	1.00			59.00
12142881	198						1.53	310.94	0.99	1.00			59.00
12142882	198						1.53	310.94	0.99	1.00			59.00
12142883	198						1.53	310.94	0.99	1.00			59.00
12142884	198						1.53	310.94	0.99	1.00			59.00
12142885	198						1.53	310.94	0.99	1.00			59.00
12142886	198						1.53	310.94	0.99	1.00			59.00
12142887	198						1.53	310.94	0.99	1.00			59.00
12142888	198						1.53	310.94	0.99	1.00			59.00
12142889	198						1.53	310.94	0.99	1.00			59.00
12142890	198						1.53	310.94	0.99	1.00			59.00
12142891	198						1.53	310.94	0.99	1.00			59.00
12142892	198						1.53	310.94	0.99	1.00			59.00
12142893	198						1.53	310.94	0.99	1.00			59.00
12142894	198						1.53	310.94	0.99	1.00			59.00
12142895	198						1.53	310.94	0.99	1.00			59.00
12142896	198						1.53	310.94	0.99	1.00			59.00
12142897	198						1.53	310.94	0.99	1.00			59.00
12142898	198						1.53	310.94	0.99	1.00			59.00
12142899	198						1.53	310.94	0.99	1.00			59.00
12142900	198						1.53	310.94	0.99	1.00			59.00
12142901	198						1.53	310.94	0.99	1.00			59.00
12142902	198						1.53	310.94	0.99	1.00			59.00
12142903	198						1.53	310.94	0.99	1.00			59.00
12142904	198						1.53	310.94	0.99	1.00			59.00
12142905	198						1.53	310.94	0.99	1.00			59.00
12142906	198						1.53	310.94	0.99	1.00			59.00
12142907	198						1.53	310.94	0.99	1.00			59.00
12142908	198						1.53	310.94	0.99	1.00			59.00
12142909	198						1.53	310.94	0.99	1.00			59.00
12142910	198						1.53	310.94	0.99	1.00			59.00
12142911	198						1.53	310.94	0.99	1.00			59.00
12142912	198						1.53	310.94	0.99	1.00			59.00
12142913	198						1.53	310.94	0.99	1.00			59.00
12142914	198						1.53	310.94	0.99	1.00			59.00
12142915	198						1.53	310.94	0.99	1.00			59.00
12142916	198						1.53	310.94	0.99	1.00			59.00
12142917	198						1.53	310.94	0.99	1.00			59.00
12142918	198						1.53	310.94	0.99	1.00			59.00
12142919	198						1.53	310.94	0.99	1.00			59.00
12142920	198						1.53	310.94	0.99	1.00			59.00
12142921	198						1.53	310.94	0.99	1.00			59.00
12142922	198						1.53	310.94	0.99	1.00			59.00
12142923	198						1.53	310.94	0.99	1.00			59.00
12142924	198						1.53	310.94	0.99	1.00			59.00
12142925	198						1.53	310.94	0.99	1.00			59.00
12142926	198						1.53	310.94	0.99	1.00			59.00
12142927	198						1.53	310.94	0.99	1.00			59.00
12142928	198						1.53	310.94	0.99	1.00			59.00
12142929	198						1.53	310.94	0.99	1.00			59.00
12142930	198						1.53	310.94	0.99	1.00			59.00
12142931	198						1.53	310.94	0.99	1.00			59.00
12142932	198						1.53	310.94	0.99	1.00			59.00
12142933	198						1.53	310.94	0.99	1.00			59.00
12142934	198						1.53	310.94	0.99	1.00			59.00
12142935	198						1.53	310.94	0.99	1.00			59.00
12142936	198						1.53	310.94	0.99	1.00			59.00
12142937	198						1.53	310.94	0.99	1.00			59.00
12142938	198						1.53	310.94	0.99	1.00			59.00
12142939	198						1.53	310.94	0.99	1.00			59.00
12142940	198						1.53	310.94	0.99	1.00			59.00
12142941	198						1.53	310.94	0.99	1.00			59.00
12142942	198						1.53	310.94	0.99	1.00			59.00
12142943	198						1.53	310.94	0.99	1.00			59.00
12142944	198						1.53	310.94	0.99	1.00			

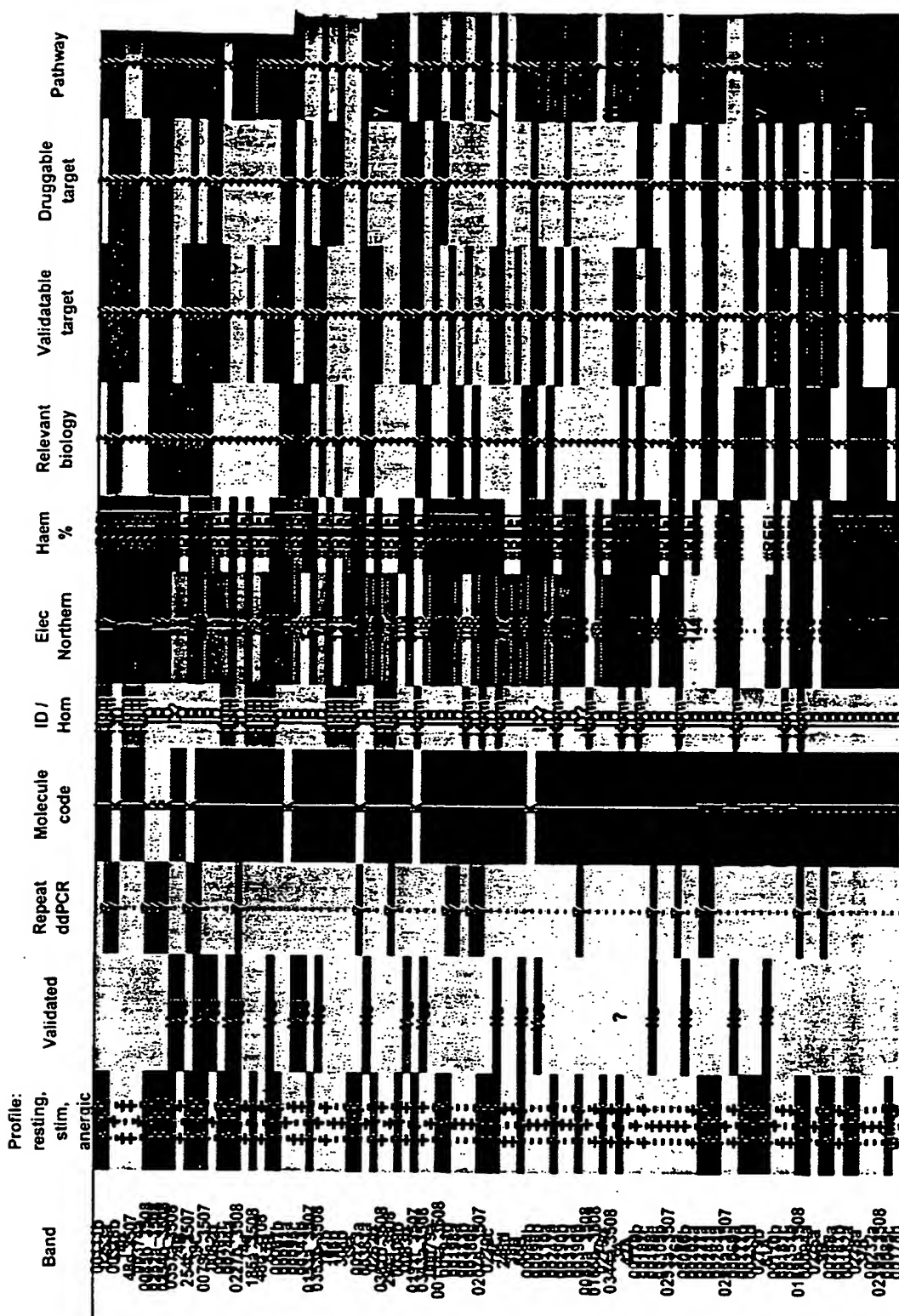


Fig. 20 Target Protein Candidates

project name	most important factor scored by manager 1	most important factor scored by manager 2	most important factor scored by manager 3	less important factor scored by manager 1	less important factor scored by manager 2	less important factor scored by manager 3
Proj 18						
Proj 05						
Proj 04						
Proj 14						
Proj 20						
Proj 13						
Proj 09						
Proj 12						
Proj 15						
Proj 01						
Proj 08						
Proj 06						
Proj 17						
Proj 19						
Proj 07						
Proj 03						
Proj 10						
Proj 16						
Proj 02						
Proj 11						

Fig. 21

Company	Disease 1	Disease 2	Disease 3	Disease 4	Disease 5	Disease 6	Disease 7	Disease 8	Disease 9	Disease 10	Disease 11	Disease 12
Company 1	Phase I		Phase II		LO/DE	LI/LO	LI/LO	LO/DE	Phase I	LI/LO	Phase II	Phase I
Company 2	Phase I	LI	Phase II			LI/LO	LI/LO			LI/LO		
Company 3	LI/LO				LI/LO	Phase II	LO	Phase II			LO	
Company 4	LO/DE				LO/Ph I	Phase I	LO/DE			LO		TS
Company 5	LI/LO				LO	Phase I						
Company 6	LO/DE							Phase I	LO/DE	LI/LO	Phase I	LI
Company 7	LI/LO	Phase II							LO		LO	LI/LO
Company 8	LO/DE					Phase I	LI/LO	LI/LO	LI/LO	Phase II		
Company 9	LI/LO				Phase III							
Company 10					LO							
Company 11	LO				LO							LI/LO
Company 12	LO		Phase II		LO		Phase III					LI
Company 13	LI		LI/LO		LI/LO			LI/LO			LI/LO	
Company 14	LI/LO		LI		LI					LI		
Company 15	LI		LI		LI		Phase II					
Company 16	LI				LI/LO	LI/LO						
Company 17	Phase II							Phase I				
Company 18			Phase II									
Company 19	LI		LI				LI					
Company 20				LI/LO								
Company 21	LI/LO				Phase III					TS		
Company 22	TS					TS						
Company 23	LI		TS		LO/DE							
Company 24	LI											
Company 25	LI/LO											

Fig. 22



Fig. 23

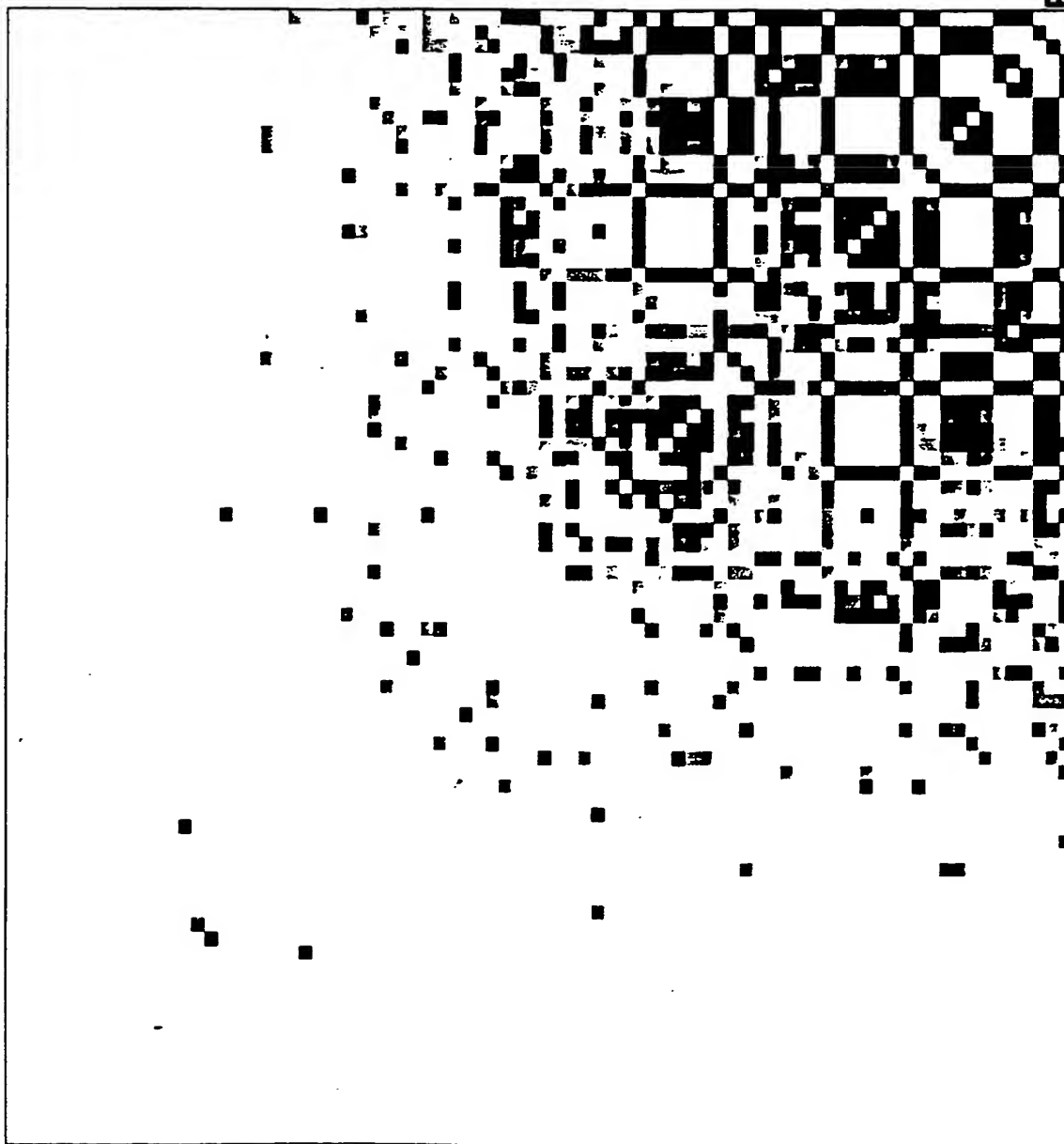


Fig. 24

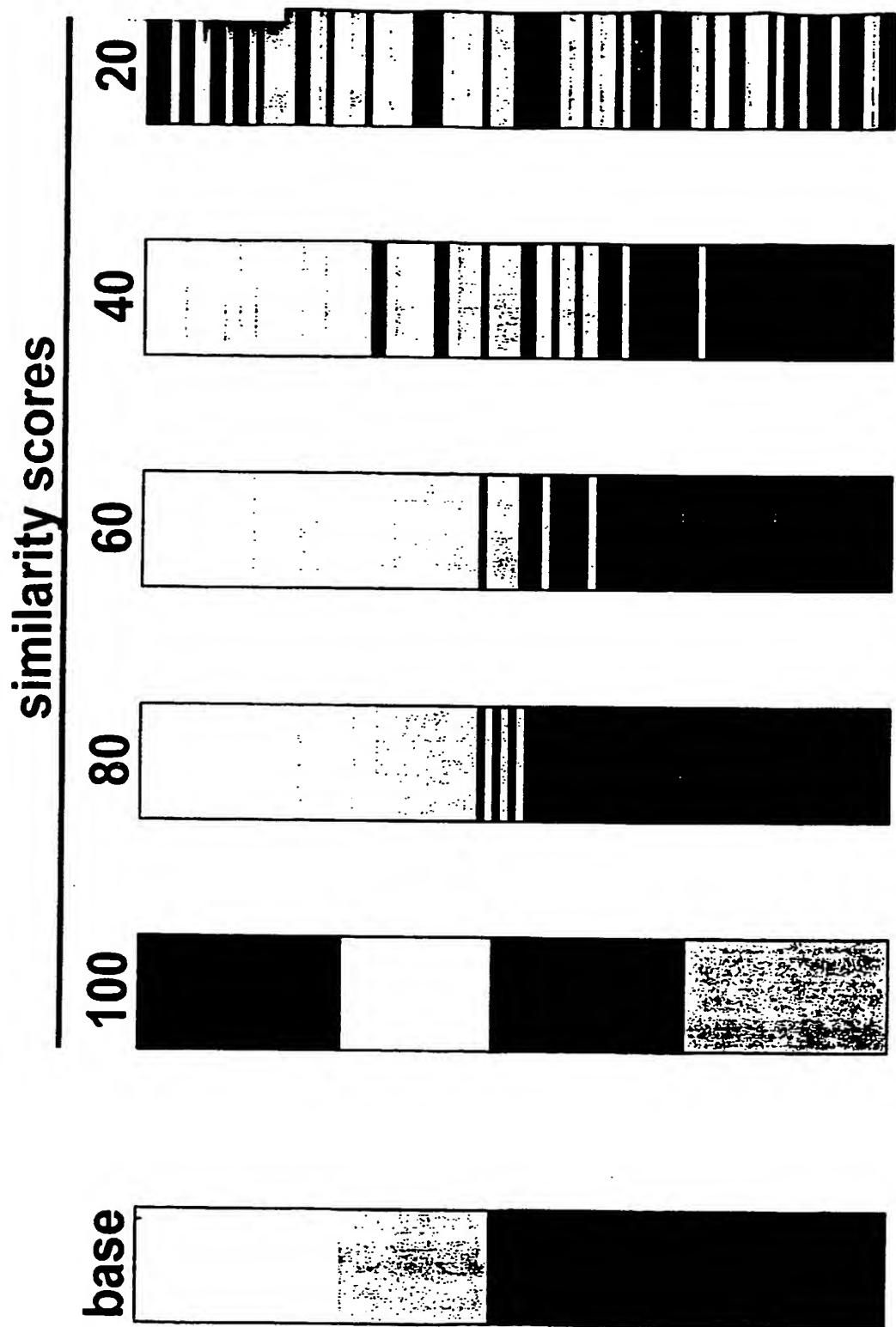


Fig. 25

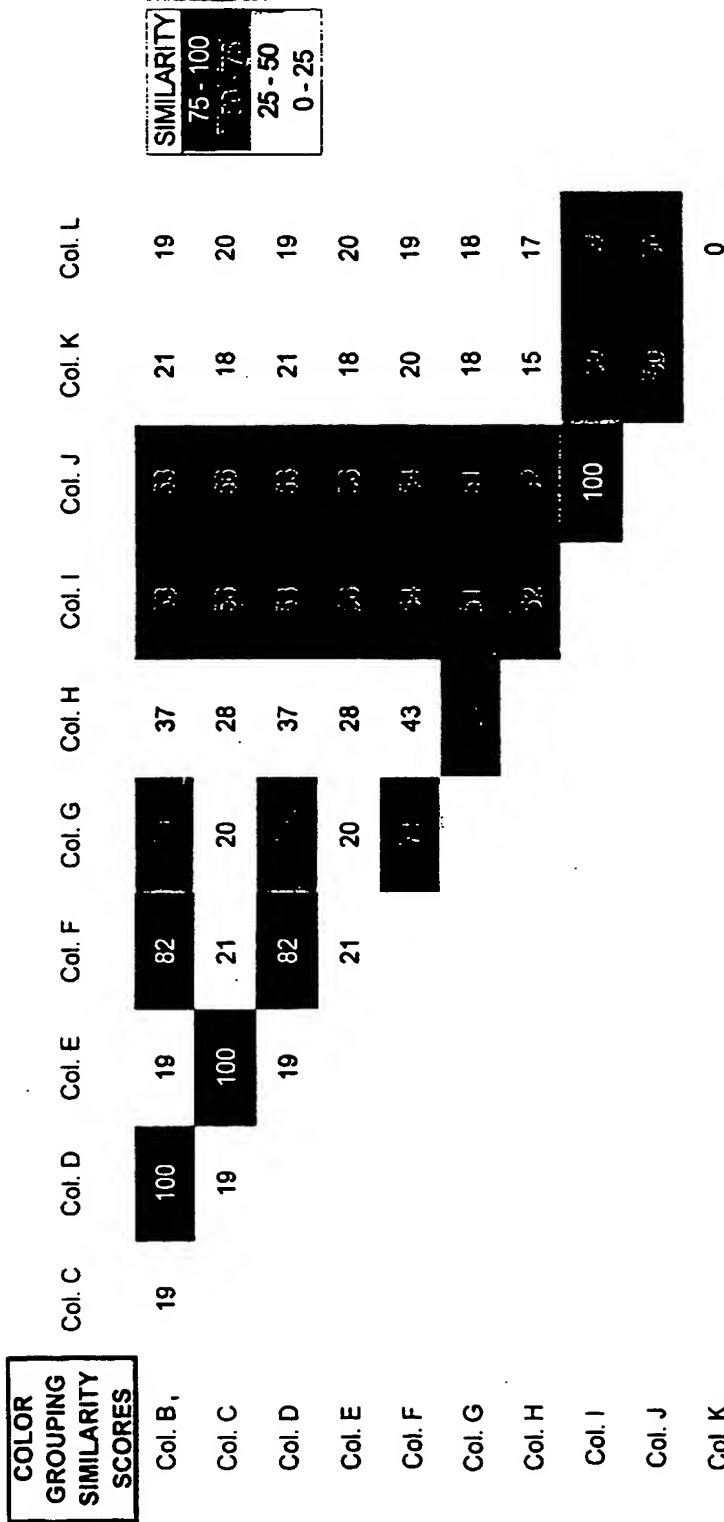


Fig. 26

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
1 February 2001 (01.02.2001)

PCT

(10) International Publication Number
WO 01/08039 A3

(51) International Patent Classification⁷: **G06K 9/00**

(21) International Application Number: **PCT/US00/20401**

(22) International Filing Date: **27 July 2000 (27.07.2000)**

(25) Filing Language: **English**

(26) Publication Language: **English**

(30) Priority Data:
09/361,122 **27 July 1999 (27.07.1999)** **US**

(71) Applicant (for all designated States except US): **ZENECA LIMITED (GB/GB); 15 Stanhope Gate, London W1Y 6LN (GB).**

(72) Inventor; and

(75) Inventor/Applicant (for US only): **LERMAN, Charles, L. [US/US]; 501 Bishop Hollow Road, Newton Square, PA 19073-3138 (US).**

(74) Agents: **BIRD, Donald, J. et al.; Pillsbury Madison & Sutro, LLP, 1100 New York Avenue, N.W., Washington, DC 20005 (US).**

(81) Designated States (national): **AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.**

(84) Designated States (regional): **ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).**

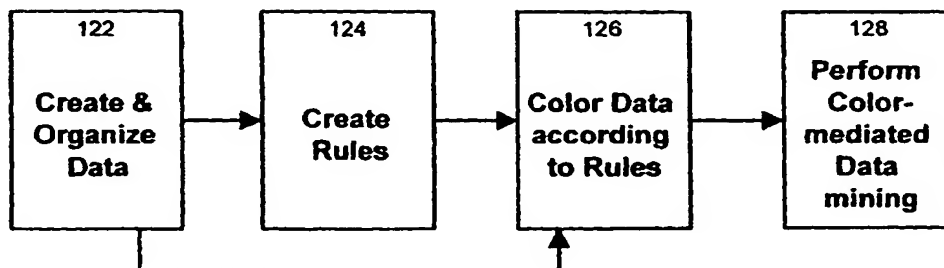
Published:

- *With international search report.*
- *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.*

(88) Date of publication of the international search report:
22 March 2001

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: **ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL DATA SETS USING LOW-RESOLUTION DATA GROUPING**



(57) Abstract: Methods, systems and devices for operating on data provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and apply at least one of the grouping rules to the data. The data may be in a table, wherein the at least one grouping rule applies to at least one user-selectable column of the table. The grouping rule defines breakpoints corresponding to the user-definable number of groups, and application of the at least one rule to the data divides the data into groups based on the breakpoints. The grouped data is presented in a manner that visually distinguishes the groups, sometimes by coloring an aspect of the data according to the rules.

WO 01/08039 A3

INTERNATIONAL SEARCH REPORT

Inter. Application No
PCT/US 00/20401

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06K9/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, COMPENDEX, INSPEC, WPI Data, IBM-TDB, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>ANONYMOUS: "Dynamic Layout Mechanism for the Massive-Node Server Status Monitor"</p> <p>IBM TECHNICAL DISCLOSURE BULLETIN, vol. 36, no. 5, 1 May 1993 (1993-05-01), pages 169-170, XP000408951</p> <p>New York, US</p> <p>the whole document</p> <p style="text-align: center;">--- -/--</p>	1-48

☒ Further documents are listed in the continuation of box C.

☐ Patent family members are listed in annex.

* Special categories of cited documents:

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

X document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

Y document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

G document member of the same patent family

Date of the actual completion of the international search

12 January 2001

Date of mailing of the international search report

22/01/2001

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Granger, B

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 00/20401

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>EISEN M B ET AL: "Cluster analysis and display of genome-wide expression patterns" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, US, NATIONAL ACADEMY OF SCIENCE. WASHINGTON, vol. 95, December 1998 (1998-12), pages 14863-14868, XP002140966 ISSN: 0027-8424 page 14863, right-hand column, paragraph 3; figure 2</p> <p>---</p>	1-48
A	<p>STANTON D T ET AL: "Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery" JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, JAN.-FEB. 1999, ACS, USA, vol. 39, no. 1, pages 21-27, XP000971515 ISSN: 0095-2338 the whole document</p> <p>-----</p>	1-48

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
1 February 2001 (01.02.2001)

PCT

(10) International Publication Number
WO 01/008039 A3

(51) International Patent Classification⁷: G06K 9/00

(21) International Application Number: PCT/US00/20401

(22) International Filing Date: 27 July 2000 (27.07.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/361,122 27 July 1999 (27.07.1999) US

DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

(88) Date of publication of the international search report:
22 March 2001

(72) Inventor; and

(75) Inventor/Applicant (*for US only*): LERMAN, Charles, L. [US/US]; 501 Bishop Hollow Road, Newton Square, PA 19073-3138 (US).

(48) Date of publication of this corrected version:
8 August 2002

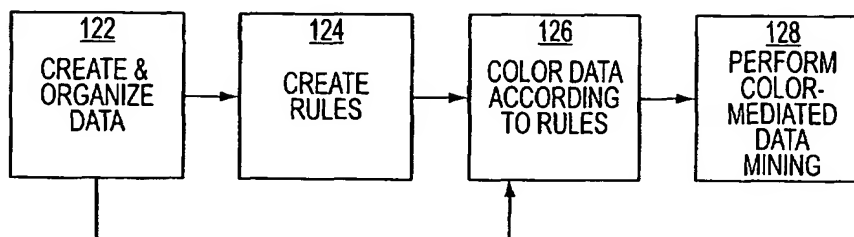
(74) Agents: BIRD, Donald, J. et al.; Pillsbury Winthrop LLP, 1600 Tysons Boulevard, McLean, VA 22102 (US).

(15) Information about Correction:
see PCT Gazette No. 32/2002 of 8 August 2002, Section II

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ,

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL DATA SETS USING LOW-RESOLUTION DATA GROUPING



(57) Abstract: Methods, systems and devices for operating on data provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and apply at least one of the grouping rules to the data. The data may be in a table, wherein the at least one grouping rule applies to at least one user-selectable column of the table. The grouping rule defines breakpoints corresponding to the user-definable number of groups, and application of the at least one rule to the data divides the data into groups based on the breakpoints. The grouped data is presented in a manner that visually distinguishes the groups, sometimes by coloring an aspect of the data according to the rules.

WO 01/008039 A3

**ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL
DATA SETS USING LOW-RESOLUTION DATA GROUPING**

A portion of the disclosure of this patent document contains material
5 which is subject to copyright protection. The copyright owner has no objection to
the facsimile reproduction by anyone of the patent document or the patent
disclosure, as it appears in the Patent and Trademark Office patent file or records,
but otherwise reserves all copyright rights whatsoever.

10 **1. Field of the Invention**

This invention relates to analysis and pattern recognition of data. More
particularly, this invention relates to methods, systems and devices and
combinations thereof for analysis and pattern recognition in large sets of
multidimensional data using low-resolution data grouping.

15 **2. Background**

With the advent of computerization and the low cost of data storage and
acquisition, people in many endeavors are now accumulating very large sets of
data. For example, scientists in drug and chemical companies now use automation
to perform so-called high-throughput screening ("HTS") of chemical compounds.
20 HTS uses automated, relatively low-cost techniques to obtain various items of
information about chemical compounds. The goal of using HTS is to obtain
information about a very large number of compounds in a quick and relatively
low-cost manner. Having accumulated a very large HTS data set, it is necessary
to evaluate the data in order to determine which, if any, of the analyzed

compounds warrants further investigation. However, the results of such HTS tend to be very large sets of multidimensional data, on the order of thousands of rows and dozens of columns, and so it is very difficult to make decisions just by looking at the data.

5 In addition to the very large amounts of data produced by HTS, difficulties in existing data handling and analysis methods include the following:

- Data comes from very diverse sources, including HTS laboratories, physical measurements, bio-scientists' laboratories, various computational
10 software programs, etc., and the different sources tend to have very diverse kinds of output including numbers, text, mixed data types, error notations, blank data, replicate data (more than one value per compound), etc.
- Not all sources produce data on the same list of compounds, or in the same order.
- 15 • Some data values are misleadingly too precise, i.e., have high relative experimental errors or noise, and can easily be over-interpreted.
- Medicinal chemists have to weigh very different kinds of factors (for example, molecular weight vs. dose-responsiveness vs. ClogP vs. secondary biology vs. selectivity across screens) in trying to determine
20 which are the best compounds or clusters of compounds to which to devote further work.

SUMMARY OF THE INVENTION

This invention solves the above and other problems by providing automated tools to help with and speed up these data handling and analysis processes. These tools embody some assumptions about how the data should be
5 treated by internalizing the most generally acceptable assumptions, but leaving more idiosyncratic decisions to individual users.

A central concept on which this invention is based is grouping data into a relatively small number of categories using low-resolution data grouping. The grouping is visualized by assigning colors to data groups, e.g., in spreadsheets.
10 Grouping of data potentially changes the precision of the data.

This categorization of data has several major benefits, including:

- creating a visual means of finding data patterns;
- beneficially blurring small variations in numerical data that are, in practice, excessively fine distinctions, possibly due to experimental
15 error; and
- providing, in the colors themselves, a means or “common currency” to evaluate candidates across a wide range of data types.

Accordingly, in one aspect, this invention provides mechanisms to expedite pattern recognition in large sets of multidimensional data, such as those
20 that chemists assemble when evaluating hits from high-throughput screening (HTS) and deciding which ones will get priority for further investigation. In controlled trials, this invention has reduced the time to evaluate real data sets, from days of intense human effort, which is vulnerable to errors due to volume or fatigue, to a few minutes of automation with graphical presentation of results.

It quickly becomes obvious upon using the system that the tools also have value in data-handling areas other than HTS. Examples include selection and management of any kind of tabulated data, e.g., portfolio management for any kind of rated portfolios, selection of drug candidate compounds, selection and management of proteins that are candidates for targets for drugs, selection and management of research projects competing for resources, and evaluating employee performance or job candidates.

The system of this invention includes a new special command menu, a set of graphical user interface worksheets, and action buttons to facilitate the coloring and color analysis processes for the user. While the central process is the data grouping and coloring, there are also new tools for the upstream, or pre-grouping and coloring processes of importing, assembling, regularizing, and characterizing data in a spreadsheet, and for the downstream processes of visualizing, scoring, comparing, and sorting large amounts of color-coded data. The data-grouping and spreadsheet-coloring tool is presently implemented with a flexible, powerful, and convenient user interface that does not require knowledge of spreadsheet macros or of the Visual Basic language (used for the system's implementation).

Accordingly, this invention provides methods, systems and devices for operating on data.

In one aspect, the method of this invention provides at least one user-defined grouping rule for grouping the data into a user-definable number of groups. At least one of the grouping rules is applied to the data. The data may be provided in a table and the grouping rule applies to at least one user-selectable column of the table. In some embodiments, the grouping rule defines breakpoints corresponding to the user-definable number of groups. Application of the rule the

data divides the data into groups based on the breakpoints. The method may include presenting the grouped data in a manner that visually distinguishes the groups. In some embodiments, the grouping rules associate colors with groups and the grouped data is presented with an aspect of the data colored according to the rules.

Sometimes the data are in labeled columns in a spreadsheet, and the grouping rule specifies at least one breakpoint and a corresponding color for each range defined by the breakpoint. The grouped data are presented by coloring each data item in one labeled column of the data based on the breakpoint and the corresponding color of the breakpoint.

The breakpoints may be numeric or textual values. In some embodiments, the breakpoint is determined automatically based on the data.

Sometimes the data are provided in a table, and backgrounds of table cells are colored according to the rules.

The number of groups may be fewer than a number of possible data values.

In another aspect, this invention is a method of operating on data by providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups. The grouping rule is applied to the data to generate grouped data. At least one user-defined scoring rule is used to score grouped data according to user-defined scores. The scoring rule is applied to the grouped data to score the grouped data.

In yet another aspect, this invention is a method of operating on data, in which data are grouped by applying to the data at least one user-defined grouping rule for grouping the data into a user-definable number of groups. The grouped

data are scored by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

In some embodiments the data can be a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be
5 used in scoring the cases. The scoring applies the function to the data to obtain a score for each case. Sometimes the method includes sorting the scored cases by score, individually or by cluster, as described below.

The notion of clustering is that subsets of the various cases may be
10 associated into clusters by having identical entries in any user-selected column of data, known as a clustering column. In some embodiments of the invention, the integrated clusters are treated by averaging the properties of all the cases which comprise each cluster.

Thus, according to aspects of this invention, in order to facilitate analysis
15 and pattern recognition in large, multidimensional data sets, the precision of the data is potentially changed (implemented, e.g., by grouping the data) and then the data are presented for visualization (implemented, e.g., by coloring the data).

BRIEF DESCRIPTION OF THE DRAWINGS

This file contains at least one drawing executed in color. Copies of this
20 patent with color drawings will be provided by the United States Patent and Trademark Office upon request and payment of the necessary fee.

The above and other objects and advantages of the invention will be apparent upon consideration of the following detailed description, taken in

conjunction with the accompanying drawings, in which the reference characters refer to like parts throughout and in which:

FIGURE 1 shows a typical computer system on which the present invention operates;

5 **FIGURE 2** shows an overview of the functionality of the present invention;

FIGURES 3A-3B depict a display of data in a spreadsheet;

FIGURES 4A-4B show a color control rules worksheet according to one embodiment of the present invention;

FIGURES 5A-5B show data coloring rules;

10 **FIGURES 6A-6C** show a data coloring control panel and a flow chart of the data coloring process, respectively;

FIGURES 7A-8B show the worksheet of **FIGURE 3A** and **3B** after various coloring rules in **FIGURE 4A** have been applied;

FIGURES 9A, 9B, 10A, and 10B depict displays of data in spreadsheets;

15 **FIGURES 11A and 11B** show the form of the cluster control worksheet according to one embodiment of the present invention;

FIGURES 11C-11D shows control panels from the cluster control worksheet of **FIGURES 11A-11B**;

20 **FIGURE 12** shows the enlarging of the cluster starts mechanism according to one embodiment of the present invention;

FIGURES 13A-13D show the application of vertical display re-scaling according to one embodiment of the present invention;

FIGURES 14A-14D and 15A-15B show the application of the scoring and sorting of clusters according to one embodiment of the present invention;

FIGURES 16A-16N, 16P and 16Q show aspects of the application of the dose-response scoring and estimation of potencies according to one embodiment of the present invention;

FIGURES 17A-17B show the application of the sheet statistics tool
5 according to one embodiment of the present invention;

FIGURES 18A-18D show the application of the scoring and sorting of clusters for the purpose of project prioritization and management according to one embodiment of the present invention;

FIGURES 19-24 show examples of the application of this invention to
10 various types of data; and

FIGURES 25 and 26 show application of an aspect of this invention.

DETAILED DESCRIPTION OF THE PRESENTLY PREFERRED EXEMPLARY EMBODIMENTS

Overview

15 FIGURE 1 shows a typical computer system 100 on which the present invention operates. The computer system 100 includes a processor (CPU) 102 connected to a memory system 104 and a display 106. The computer system also includes various input devices including a keyboard 108 and a mouse 110 or other pointing device. Internal storage 112 (e.g., a hard disk, a CD ROM and the like)
20 and external storage 114 (such as a floppy disk, CD ROM and the like) are also provided.

Various aspects of this invention are implemented as computer software programs or algorithms 116 which run on the computer system 100. The software programs 116 can reside in the internal storage 112, the external storage 114,

and/or in the memory 104. The software programs 116 operate on data 118 which is provided, e.g., on the external storage 114. The software programs 116 operate in a standard and known manner by being executed on the processor 102 of the computer system 100.

5 In some embodiments of the present invention, the user can create and modify various executable rules 120 which can operate on the data 118. For the sake only of explanation, the rules 120 are depicted separately from the data in the figures. However, as explained in more detail below, some or all of the rules 120 can be part of the data 118.

10 In preferred embodiments, the computer system 100 is capable of running the spreadsheet program EXCEL™ 95 (hereinafter "EXCEL") from Microsoft Corporation, and the software computer programs 116 are written in Microsoft Corporation's Visual Basic (hereinafter "VB") and are provided as an add-in to EXCEL. A single copy of software thus serves all data files on a particular
15 machine. To conserve EXCEL resources, in some embodiments, the package self-installs the add-in when the user opens a data file, and un-installs the add-in when the last data file in memory is closed.

 In a preferred embodiment, this invention works entirely within the environment of EXCEL. EXCEL structures data files as workbook files which
20 contain worksheets. The programs 116 of this invention consist of special EXCEL worksheets, called control sheets, on which input data is written by the user into designated labeled cells. The control sheets are part of the same EXCEL workbook file as the data. The control sheets also contain action buttons to execute the various procedures associated with this invention. The rules 120 are formed by
25 setting various parameters in the control sheets.

When the workbook file is saved, the parameters (for the rules 120) are stored on the control sheets along with the data, and they can be modified and/or re-executed at any time without having to re-enter anything. The results of operations are automatically written as worksheets in the same workbook file, providing a convenient, integrated data environment in a single file.

The system according to the present invention operates, in one aspect, in accordance with FIGURE 2. Recall that the user's aim is to perform analysis and pattern recognition in large, multidimensional data sets using (potentially low resolution) data grouping. To this end, the user and/or the system will create rules for coloring and presenting the data. First (at 122) a user creates and organizes the data 118. Various tools (discussed below) are provided to aid in the creation and organization of the data. Then (at 124) the user creates rules 120 for operating on the data 118. The rules 120 can be created before or after the data 118, rules can be reused for different sets of data and multiple rules can apply to the same data. The creation and operation of rules are discussed in greater detail below. Once the data 118 and the rules 120 are created, the user then selects some (or all) of the rules to apply to the data (at 126). Specifically, the user groups and thereby colors the data according to selected rules. With the data grouped and colored according to the rules, the user can then perform group/color-mediated data mining (at 128).

FIGURES 3A-3B show views of the program of this invention in operation with a sample EXCEL sheet 300, denoted "DEMO 1" (302) containing data (not all the data in the sheet is visible). The views of EXCEL worksheets shown in the various figures and examples that follow are the views that are presented on the display 106 of the computer system 100. Sheets in an EXCEL workbook are labeled with tabs at the bottom of the worksheet. The data on the "DEMO1" sheet

300 consists of eight columns of data for each of a number of compounds. The compounds are denoted "Cmpdxx", where "xx" ranges from "01" to the number of compounds. In FIGURE 3B, the last compound visible on the data sheet is "Cmpd58". The eight columns are headed:

- 5 1. "Cmpd" (column A);
2. "Series" (column B);
3. "Test1" (column C);
4. "Test2" (column D);
5. "Test3" (column E);
- 10 6. "HTS SPA Dose-Resp % Inhib @3x10-6M" (column F);
7. "HTS SPA Dose-Resp % Inhib @ 1x10-6M" (column G);
8. "HTS SPA Dose-Resp % Inhib @ 3x10-7M" (column H); and
9. "HTS SPA Dose-Resp % Inhib @ 1x10-7M" (column I).

In addition to the "DEMO 1" worksheet 300, the EXCEL workbook shown

15 in FIGURES 3A and 3B has seven other worksheets, denoted "DEMO 2" 304; "DEMO 3" 306; "clusterinfo DEMO" 308; "Append Control" 310; "Color Control" 312 and "Cluster Control" 314. The last three worksheets, denoted respectively "Append Control"; "Color Control" and "Cluster Control," contain various rules and controls (to be discussed below). The data in worksheets

20 denoted "DEMO 1" 302; "DEMO 2" 304; "DEMO 3" 306; and "clusterinfo DEMO" 308 correspond to data 118 (FIGURE 1) and the controls or rules in the worksheets denoted "Append Control" 310; "Color Control" 312, and "Cluster Control" correspond to the rules 120 (FIGURE 1).

FIGURES 4A-4B show a color control rules worksheet (312, denoted "color control") according to the present invention, as displayed on display 106 of the

25

computer system 100. The color control worksheet 312 is shown with some rules already in place, i.e., having values set, and other rules left blank. These rules are shown as examples only, and, as with any of the other types of rules, any or all of the rules can be set by the user. A typical data coloring rule 130 is shown in

5 **FIGURE 5A.** The rule 130 has already been set up and operates on the appropriate data when selected by a user (using mouse 110, **FIGURE 1** or some other pointing device) in the area 132 marked "Click here to run these". The rule 130 (as with all of the preferred color control rules) has four parts, namely the name of the sheet 134 containing the data on which the rule is to operate ("DEMO 1" in the example

10 of **FIGURE 5A**); the columns 136 of data of the sheet on which the rule is to operate ("E" in the example of **FIGURE 5A**); the number of colors 138 to be used by the rule 130; a number of breakpoints 140 (denoted "break 1" to "break 4" in the example of **FIGURE 5A**); and a corresponding number of colors 142 for each range defined by the breakpoints (denoted "color 1" to color 4" in the example of

15 **FIGURE 5A**). Actually, as explained below, the number of breakpoints is one less than the number of colors.. In the specific example shown in **FIGURE 5A**, the rule has three breakpoints of 1, 5 and 10, defining four ranges with four corresponding colors 142, namely light green, yellow, orange and red. Preferably the named colors are also depicted in the actual colors, so that, in this example, the

20 background of the word "lightgreen" is shown in light green, the background of the word "yellow" is shown in yellow and so on.

In this invention it is preferable to show data and meta-data (headings etc.) in color. In some embodiments, the coloring is implemented by showing a background area of the text representing the data in the appropriate color.

25 Sometimes the actual text representing the data is shown in the appropriate color.

In presently preferred embodiments, the font color is only changed in cases where necessary to improve contrast with the background color for readability. Only two font colors, dark (black) and light (pale gray), are used in the presently preferred embodiment. Combinations of both approaches can be used. For example, the background section of the word "yellow" is preferably shown in the color yellow. It is also possible to show the word itself, i.e., the font, in the color yellow, as long as that color is distinguishable from the background.

The particular rule 130 shown in FIGURES 4A and 5A, operates as follows, when selected:

In sheet "DEMO 1" 302, in column E, values less than or equal to 1 (break 1) are colored light green (color 1); values in the range 1 to 5 (between break 1 and break 2) are colored yellow (color 2); values in the range 5 to 10 (break 2 to break 3) are colored orange (color 3); and values greater than 10 (break 3) are colored red (color 4).

Another typical data coloring rule 130-1 from the color control sheet 312 is shown in FIGURE 5B. The rule 130-1 is set up to operate on columns "C" and "D" of sheet "DEMO 1". The rule 130-1 uses three (3) breakpoints (break1=0.1, break2=1 and break3=5) defining four ranges with four (4) corresponding colors ("lightgreen", "yellow", "orange", and "red").

The results of applying the rule 130-1 of FIGURE 5B to the data in sheet "DEMO 1" (302, FIGURE 3) are shown in FIGURES 7A-7B. As can be seen from FIGURES 7A-7B, after application of the rule 130-1, all of the data in columns C and D of the sheet labeled "DEMO 1" has been colored according to the rule. Specifically, data having a value less than or equal to break 1 (0.1) have been colored light green; data values in the range between break 1 and break 2 (0.1 to

1) have been colored yellow; data values in the range between break 2 and break 3 (1 to 5) have been colored orange; and data values greater than break 3 (5) have been colored red.

The results of applying all of the other color control rules shown in **FIGURES 4A-4B** to the data in sheet "DEMO 1" are shown in **FIGURES 8A-8B**.
 The rules can be applied individually (as shown above with respect to **FIGURES 7A-7B**), or they can be all be applied at the same time. In order to apply all rules to a particular data set (sheet), each rule can be individually selected or the area labeled "RE-RUN ALL RULES FOR SHEET NAMED DEMO 1" (on the right side of **FIGURE 4A**) can be selected. Note that if two rules apply to the same column of the same sheet, the second rule run on that column will override the first rule run on that column.

To create a coloring rule a user performs the following (with reference to **FIGURE 6B**):

- (1) Select the "COLOR CONTROL" sheet **312** and pick a control panel on that sheet to use (an empty panel or one containing a rule no longer needed) (at **600**). All control panels on a sheet can be cleared by clicking the button labeled "CLEAR ALL ENTRIES ON THIS SHEET" (**318** in **FIGURE 4A**).
- (2) In the selected control panel, enter the name of the sheet to be colored (at **602**).
- (3) In the selected control panel, enter a column or columns (at **604**).

For multiple columns, either list them separated by commas, or use a colon or hyphen to denote ranges, or some combination. For example, "A:D,F" means columns A,B,C,D, and F. To aid in

choosing columns, the user can right-click on the cell containing the name of the data sheet, and pick "Open Twin Screen" from the shortcut menu that appears, to create a special dual display. This also creates a "Close Twin Screen" button to go back.

5 (4) Choose a number of colors to use (at 606), either by entering the number of colors or by repeatedly clicking the gray button adjacent the cell labeled "# of colors". In preferred embodiments, the system allows for five breakpoints and six colors per rule. Accordingly, the numbers will cycle from 1 to 6, and various cells
10 below them will be blacked out accordingly.

(5) Enter the breakpoints that define the color groups (at 608), in any of three modes:

a) Numeric data, manual mode: enter numbers to form the breakpoints, i.e., the boundaries between the
15 color groups, one less than the number of colors, in increasing numerical order. Cells whose values exactly equal a breakpoint value will be colored with the lower group (breakpoint 1 is colored with color 1, etc.)

20 b) Numeric data, automatic mode: enter either "value", "log", or "count" as the first breakpoint. If multiple columns have been chosen, the user must also enter "yes" or "no" opposite "Re-scale all?" at the bottom of the panel, to indicate whether each column
25 should get its own auto-breakpoints, or whether the

auto-breakpoints of the first column (first in list in the rule, not first on the data sheet) should be used for all.

This mode reports information about the breakpoints it determines, and thus could also be used to explore the distribution of numerical values in a column prior to a final *manual* breakpoint selection.

- c) Text data: enter the strings to be matched and colored, in preferred embodiments, up to five (5) in number. Matching is case-insensitive unless the string is enclosed in double quotes (“ and ”); otherwise, no quotation marks are necessary.

Several special text strings act as operators if entered as the first word in a rule cell:

<u>Rule Entry</u> (OPERATORS need not be uppercase—here only for emphasis)	<u>Meaning</u>
test string	color data cell if its whole content matches the test string
NOT test string	<ul style="list-style-type: none"> • color data cell if its whole content does not match the test string; • will not color numeric cells
CONTAINS test string	color data cell if contains the test string as a substring anywhere
NOTCONTAINS test string	<ul style="list-style-type: none"> • color data cell if it does not contain the test string anywhere; • will not color numeric cells

<u>Rule Entry</u>	<u>Meaning</u>
(OPERATORS need not be uppercase—here only for emphasis)	
BEGINS test string	color data cell if it begins with the test string
ENDS test string	color data cell if it ends with the test string
* (an asterisk)	(wildcard) color data cell containing any data, including numeric cells
BLANK	color data cells that are blank

Using quotes to force matching to be case-sensitive also works with strings that follow an operator.

It is possible to construct a text-coloring rule in which certain cells may satisfy more than one of the “breakpoint” values. For example, if a rule says that “active” is colored green and “contains act” is colored red, then the word “active” in a cell would satisfy both. In such cases, the system colors the cell according to the *first* condition satisfied on the list of breakpoints. This dependence on the order can be used advantageously to achieve complex coloring conditions. The sequence of conditions can be considered as a series of filters, through which only the as-yet-uncolored cells fall through to the next decision.

- (6) Enter the names of the colors to use (at 610), in the order corresponding to the breakpoints. A display of color samples is provided at the right side of the Color Control sheet 312. A user need only enter the name, and the appropriate cell will become colored when the tool is executed. If the user wants the color to display immediately, he can copy and paste the sample cell into the rule's color cell. A special pseudo-color named "SKIP" is used to tell the system not to color the cells whose data falls in this group.
- (7) When the rule has been created, the user executes the rule by selecting the rule's "CLICK HERE TO RUN THESE" button on the panel filled in (at 612, FIGURE 6C).
- (8) To create different coloring rules for other columns, repeat the above in additional control panels. If the user runs out of control panels, he can create more control panels by copying an existing one and pasting it onto a blank section of the color control sheet.

To the extent that a single panel cannot hold all the requirements for a particular rule, a user can combine two or more panels to create a particular rule. For example, if a user needs ten (10) breakpoints, two panels can be used.

With reference to the coloring rule is shown in FIGURE 6A, once the rule has been set, a number of parameters are stored in the system. The parameters are "sheet name" ("DEMO 3" in FIGURE 6A), column specification, number of colors, array of breakpoints, array of colors, and multicolumn scaling mode.

The data coloring mechanism operates as follows, with reference to the flowchart of **FIGURE 6C.**:

1. The user enters the parameters into a rule panel on a "COLOR CONTROL" worksheet **312**, e.g., as described above with reference to the panel of **FIGURE 6A.**
2. The user selects (clicks) the activation button (labeled "Click here to run these") on that panel (at **612**). This causes the system to:
 - (A) Read and interpret the parameters from the panel (at **614**).
The system can identify which button was clicked using the Visual Basic ("VB") "caller" property, described in more detail below. The parameters are then read based on the identity of the cell location of the button using the VB "TopLeftCell" property. The system retrieves the parameters (sheet name, column specification, number of colors, array of breakpoints, array of colors, and multicolumn scaling mode) from cells in this panel by relative reference to the button cell.
 - (B) Next, determine the mode of the coloring rule (at **616**) (i.e., numeric v. text or manual v. automatic, and, if automatic, which of value, log or count). This uses the analysis of the first breakpoint entry.
 - (C) Compile a list of the columns specified in the "column specification" parameter (at **618**). This is done by scanning the various areas contained in the selection, as follows:

```

For Each singlearea In Selection.Areas
  For Each c In singlearea.Columns
    If Not CountEmpty Then
      lr = LastRowInColumn(c.Column)
    End If
  
```

```

      If Not CountEmpty And lr = 0 Then
          'skip this empty column
          ncols = ncols - 1
      Else
          ' add this column to the list
5         icol = icol + 1
          colnumarray(icol) = c.Column
      End If
      Next c
Next singlearea

```

10

(D) If an auto-breakpoint mode is being used (determined at 620), analyze the data values to determine the breakpoints (at 622). This is done by:

- (i) Collecting statistics on the data distribution in each specified column; and
- (ii) Calculating the automatic break points for the appropriate mode. For example, the auto-value breakpoints are determined as follows:

15

```

20 If breakmode = "VALUE" Then
    interval = (maxvalue - minvalue) / ncols
    break(0) = minvalue
    For ibreak = 1 To ncols - 1
        break(ibreak) = break(ibreak - 1) + interval
25 Next ibreak

```

(iii) Displaying the results for user approval or cancellation.

(C) Loop through the cells in the chosen columns on the chosen worksheet (at 624).

30

(D) Compare each cell's value to the list of breakpoints (at 626). If the coloring rule is in text mode, use the meanings of the special breakpoint operators ("contains", "blank", asterisk "*"; or quotation marks).

5

(E) When a match is found, apply the appropriate color (at 628).

The code below illustrates the processes (D) and (E) for numeric breakpoints:

10

```

For Each cell In Range(Cells(StartColoringRow, colnum),
                      Cells(FinishColoringRow, colnum))
    cvalue = cell.Value
    colored = False
15    If IsNumeric(cell.Value) Then
        If Not IsEmpty(cell) Then
            ' (have to test both because EMPTY is numeric)
            For ibreak = 1 To ncolors - 1
                If cvalue <= break(ibreak) Then
20                    If Color(ibreak) <> "SKIP"
                        cell.Interior.ColorIndex =
                            Color(ibreak)
                        Call TextContrast(cell)
                    End If
                colored = True
25            Exit For
            End If
            Next ibreak
        If colored = False Then
30            ' Not hit yet? Must be top category, so:
            If Color(ncolors) <> "SKIP"
                cell.Interior.ColorIndex = Color(ncolors)
                Call TextContrast(cell)

```



```

        End If
        colored = True
    End If
    End If
5      Else      ' not numeric - just don't color it
End If
Next cell

```

The operation of the data coloring tool of this invention will now be
 10 described in greater detail. Each coloring rule is provided in a coloring control
 panel that has the general form of a coloring rule as shown in **FIGURE 6A**. In one
 preferred embodiment, each coloring control panel **144** is a double-outlined unit,
 sixteen (16) cells high by two (2) cells wide. As noted above, a user is provided
 with coloring control panels on the color control worksheet **312**. A user can use
 15 the coloring control panel **144** to set the sheet and column(s) on which the rule is
 to operate, the number of colors, the various break points and the colors associated
 with those breakpoints. The sheet is set by entering its name into the cell **146**
 adjacent the cell labeled "sheet". The column (or columns) on which the rule is to
 operate is (are) set by entering its (their) name in the cell **148** adjacent the cell
 20 labeled "column(s)". The number of colors is set by the user by selecting the cell
150 adjacent to the cell labeled "# of colors". Each time the cell **150** is selected it
 increases the number of colors, up to a maximum of six (6), i.e., rotating through
 the values 1 to 6. I.e., when the cell **150** shows a "6" and is selected, it reverts
 back to "1". That is, selecting the cell **150** causes the value in the cell to cycle
 25 from "1" to "6" and then back to "1".

A Visual Basic ("VB") macro function ("*CallColorColumn*") is associated
 with the top cell **152** of the control panel **144**. When the cell **152** is selected by

the user (with the mouse 110 or the like), the function associated with that cell is executed by the computer (CPU 102). In the presently preferred embodiments, the *CallColorColumn* function extracts the button name of the cell 152 and then calls a second function ("*CallColorColumnSubroutine*") with that button name as one of the parameters. The function *CallColorColumnSubroutine* takes three
5 parameters, namely *ButtonName*, *StartColoringRow*, and *FinishColoringRow*. The two parameters *StartColoringRow*, and *FinishColoringRow* are optional.

First, the function *CallColorColumnSubroutine* determines what specific values to use for the coloring by reading them from the control panel 144. Since
10 the values are all in fixed positions relative to the selected button cell 152 that initiated the call to the function *CallColorColumn*, the values can be determined once the location of that button cell 152 has been determined. This is done using the following Visual Basic code:

```
15      Sheets("Color Control").Activate
      headingrow =
          ActiveSheet.Buttons(ButtonName).TopLeftCell.Row
      headingcol =
          ActiveSheet.Buttons(ButtonName).TopLeftCell.Column
20
```

Note that if the function *CallColorColumnSubroutine* was called from another sheet (not "Color Control") then this method will not find it.

The various parameter values are then read as follows:

```
Sheet name:
25      datacol = headingcol + 1
      sheetname = Trim(Cells(headingrow + 1, datacol).Value)
```

If there is no sheet named "sheetname" an error function is called.

Generally, in preferred embodiments, a great deal of error checking takes place to ensure that the user is provided with a friendly and useable interface to the program. Most of the error checking is not mentioned in this description,
5 however, one skilled in the art would know what kinds of error checking to implement in order to provide a user-friendly working environment.

The column(s) to be colored are specified by:

Cells(headingrow + 2, datacol).Value

10 The number of colors is specified by the variable *ncolors*, where:

ncolors = Cells(headingrow + 3, datacol).Value

15 Within the function *CallColorColumnSubroutine* there are two arrays, named *break* and *color*, which are used to store the breakpoints and colors, respectively. The first breakpoint is set as follows:

break(1) = Cells(headingrow + 4, datacol).Value

20 The value of the first breakpoint is used to determine the break mode ("NUMERIC", "VALUE", "LOG", or "COUNT"). If *break*(1) (as determined above) is numeric, then the mode is set to "NUMERIC", otherwise, if *break*(1) is one of "VALUE", "LOG", or "COUNT", then the break mode is set to that mode,
25 otherwise the break mode is set to "TEXT".

Next, the function determines whether multiple columns were specified, in which case it determines whether the user selected to re-scale all the columns.

The user's re-scale selection is determined by:

```
5      rescale_all_string = Cells(headingrow +  
      15, datacol).Value
```

Now the rest of the breakpoints (if any) are read. If the break-mode is "AUTO" then the breakpoints are set as follows:

```
10      For i = 2 To lastbreaknum  
          break(i) = Cells(headingrow + 3 + i, datacol).Value
```

Various possible errors are checked for. E.g., if any breakpoints are missing (i.e., if *break(I)* is empty, the user is notified. Also, if the break mode is "NUMERIC" and non-numeric breakpoints are set, the user is notified. If
15 numeric breakpoints are not in increasing order, the user is notified. As noted above, generally, in preferred embodiments of the present invention, a great deal of error checking is performed on all user inputs to ensure that the values are correct and consistent. Most error checking is omitted from this description.

The *CallColorColumnSubroutine* maintains an array, *colorname*, which
20 maps integers to colors. In preferred embodiments, there are fifty six (56) colors available. To use the higher numbered colors, the computer's video card must be set appropriately. Using the *colorname* array, the program next associates the user provided color names with index numbers. Specifically, for each of the (up to six in a preferred embodiment) colors specified, the user specifies an actual color
25 name, denoted *cname*. This name is determined for each color by:

```
      For j = 1 To ncolors  
          cname = Cells(headingrow + 8 + j, datacol).Value
```

The interior of each color-specifying cell is then colored by the appropriate (selected) color by setting the color property (*Interior.ColorIndex*) of the cell:

```
Cells(headingrow + 8 + j, datacol).Interior.ColorIndex =
5      Color(j),
```

where the value of the variable *j* ranges from 1 to *ncolors*.

Then the cell is further processed by a function *TextContrast*.

```
Call TextContrast(Cells(headingrow + 8 + j, datacol))
10      With the parameters read and checked, the system is ready to process and
color the selected sheet (specified at cell 146 in FIGURE 6A). The selected
columns (specified in cell 148 in FIGURE 6A) in the selected sheet are processed
one-by-one by the following program code:
```

```
Call ParseInput(InString, inspecifier)
15 For Each singlearea In Range(inspecifier).Areas
    For Each c In singlearea.Columns
        colnum = c.Column
        Call ProcessOneColumn(colnum, ncolors, break,
20                               Color, breakmode,
                                rescale_all, sheetname,
                                StartColoringRow,
                                FinishColoringRow)
    Next c
Next singlearea
25
```

The processing performed by the function *ProcessOneColumn* is as follows: The column designated by *colnum* on sheet *sheetname* is to be colored according to the breakpoints in the array *break* and the colors in the array *colors*. The designated column is colored from the row corresponding to

```
30 StartColoringRow to the row corresponding to FinishColoringRow. Note that the
```

function *ProcessOneColumn* is also provided with the break mode and the variable *rescale_all*.

Function *ProcessOneColumn* first calculates the automatic breakpoints, if necessary. Note that automatic breakpoints are determined from the whole
 5 column, even if this call says to color only a limited range of rows. If the value of *breakmode* is "VALUE" or "LOG" and the value of *rescale_all* is set to "True" Or the value of the first breakpoint (*break(1)*) is set to "VALUE" or "LOG", the program calls the function *AutoValueBreakpoints* as follows:

```
10      Call AutoValueBreakpoints(colnum, colletter, ncolors,
      break, Color, breakmode, rescale_all).
```

Otherwise, if the *breakmode* is set to "COUNT" and the value of *rescale_all* is set to "True" or the first breakpoint (*break(1)*) is set to "COUNT", then the program calls the function *AutoCountBreakpoints*, as follows:

```
15      Call AutoCountBreakpoints (colnum, colletter, ncolors,
      break, Color, breakmode, rescale_all, sheetname).
```

With the breakpoints calculated, the columns are colored according to the type of breakpoints specified by the user. Specifically, when the *breakmode* is any
 20 one of "VALUE", "COUNT", "LOG", or "NUMERIC", the system executes a function *ColorNumericColumn*. On the other hand, when the *breakmode* is "TEXT", the system executes a function *ColorNumericColumn*. The VB code for this is as follows:

```
25      Select Case breakmode
          Case "VALUE", "COUNT", "LOG", "NUMERIC"
```

```
        Call ColorNumericColumn(colletter, ncolors,  
break, Color, StartColoringRow, FinishColoringRow)  
        Case "TEXT"  
            Call ColorTextColumn(colletter, ncolors, break,  
5 Color, StartColoringRow, FinishColoringRow)  
        End Select
```

Then, when the rule in control panel **144** is selected for execution, the rule is applied to the selected column(s) (denoted in cell **148**) of the named sheet (in
10 cell **146**). For each column in the named sheet, the value in each cell is compared to the various breakpoints and the cell is colored corresponding to the appropriate breakpoint.

Examples of the application of various coloring rules in the "COLOR CONTROL" worksheet of **FIGURES 4A-4B**, are shown with reference to the data in
15 worksheet "DEMO 2" (depicted in **FIGURES 9A, 9B, 10A and 10B**).

Color-Mediated Data Mining

As noted above with reference to **FIGURE 2**, once the data have been colored according to the user-selected rules (at **126**), the user can then perform color-mediated data mining (at **128**). The presently preferred embodiment of this
20 invention provides five mechanisms (each discussed below) for color-mediated data mining, namely mechanisms to:

1. enlarge/shrink cluster starts;
2. vertically re-scale the display;
3. score and sort clusters; and
- 25 4. score and sort dose-response data.

The following discussion refers to the cluster control worksheet which is shown in FIGURES 11A-11B.

1. Enlarge/Shrink Cluster Starts

5 The “Enlarge Cluster Starts” mechanism highlights the first row of each cluster in clustered data by enlarging the font of the cell containing the cluster number or label, thus enabling size reduction of the spreadsheet for the user to focus on the color patterns. When the cell height is dramatically reduced in order to see more cells on a screen or printed page, this enlargement allows the user to
10 still read the label at the beginning of each cluster. The mechanism takes user input from a *Cluster Control* worksheet. A corresponding mechanism (“SHRINK CLUSTER STARTS”) allows for undoing the enlarging. This mechanism handles cluster numbers or textual labels. Any column can be designated as the cluster labels to be processed.

15 Operation of the mechanism is as follows:

- (1) From the “CLUSTER CONTROL” sheet 314 pick a control panel to use (one which is empty or one containing inputs no longer needed). On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input
20 information for several tools.
- (2) In the blue section, enter a sheet name and the column to be considered as the cluster labels.
- (3) Click either the blue-text “Enlarge Cluster Starts” or “Shrink Cluster Starts” button.

The program code accomplishes this by scanning the column of cluster labels, identifying any entries that are different from the one immediately above, and enlarging them. Code that carries out this function is shown below:

```

5  For Each c In Range(Cells(3, colnum), Cells(lastrow,
    colnum))
        irow = c.Row - 1
        icol = c.Column
        If c.Value <> Cells(irow, icol).Value Then
10         c.Font.Size = bigfontsize
            '      Rows(Irow + 1).RowHeight = bigrowheight
        End If
    Next c

```

Example

15 An example of the application of the enlarge cluster mechanism is shown in **FIGURE 12** which shows the application of a rule (shown in the control panel **FIGURE 11C**) from the cluster control worksheet in **FIGURE 11B** to the data of worksheet "DEMO 2" as shown after coloring in **FIGURES 10A-10B**. As shown in **FIGURE 11C**, the rule is to be applied to column B of sheet "DEMO 2".

20

2. Vertical Display Re-Scaling

The vertical re-scaling mechanism operates by taking a user-provided scale factor and then changing height of data rows to facilitate visualization of large-scale color patterns. The mechanism leaves column heading heights and column widths unchanged. This makes headings remain readable and greatly simplifies
25 examining long columns of data for color patterns.

FIGURES 13A-13D show the application of the vertical display re-scale mechanism according to the present invention. FIGURES 13A-13B show some of the data in the worksheet labeled "DEMO 3" 306 (FIGURE 13A shows the first thirty eight or so elements and FIGURE 13B shows the remaining elements of that worksheet). As can be seen from the figures, the worksheet "DEMO 3" 306 has three hundred and twenty eight (328) data entries (in rows 2-329). The user can vertically scale the display by selecting "Re-scale Vertical" from the system's special menu or by pressing a particular control key sequence (e.g., "CNTL-SHIFT-V" in a preferred embodiment). This presents the user with a dialog box 318, as shown in FIGURE 13C, which asks the user to enter a scaling factor relative to the current size. The user enters a scaling factor to enlarge or reduce or restore the display. In the example shown, the user enters a scaling factor of 0.1 which produces the vertically scaled display shown in FIGURE 13D.

Vertical scaling allows a user to get an overview of the data, based on the coloring.

The portion of the program code presented below carries out the central function of the vertical display rescaling mechanism:

```

rowspec = "2:" & lastrow ' leaves the headings unchanged,
20  i.e., readable
If factor = -1 Then
    Rows(rowspec).Rows.AutoFit
Else
    For irow = 2 To lastrow
        Rows(irow).RowHeight =
25         Rows(irow).RowHeight * factor
    Next irow
End If

```

After execution of the rescaling mechanism, as can be seen in **FIGURE 13D**, the height of each row (except the heading rows) has been scaled by *factor*, 0.1 in the example shown. In this manner, all rows of the data are made
5 visible on a single page, thereby facilitating data analysis.

3. Scoring and Sorting Clusters

Scoring and sorting clusters assign numerical scores to the color patterns of individual rows or clusters of rows, thereby enabling comparison and sorting of the clusters by score.

10 The scoring and sorting mechanism accepts user's designations of colors and corresponding relative scores. It handles cluster numbers or textual labels. Any column can be designated as the cluster labels to be processed. The mechanism scores a user-selected list of columns of data, with user-assigned relative weights, which need not be equal for all columns.

15 User input is taken from a *Cluster Control* worksheet **314** (see **FIGURES 11A and 11B**), which stores any number of parameter sets, each one with a user-specified name.

The input data is automatically sorted by cluster label before starting, in order to group the clusters together in case the user has previously sorted the data
20 by some other criterion. Then scores are normalized to remove the effects of cluster size, absolute magnitude of scoring points chosen, and absolute size of weights chosen. The results are written to two new worksheets without altering the original data sheet. The first derived sheet is for the numerical scores; the second is like the original, but has the clusters sorted into descending score order,

so that the “best” are at the top, removing the need to visually scan a long colored worksheet. The derived output sheets have names that indicate their source data sheet and the name of the parameter set used for scoring. At the user’s option, the system reversibly hides the un-scored columns in the cluster-sorted output sheet,
5 focusing attention on the data that were used in scoring.

In preferred embodiments, the system detects uncolored cells in the data and offers the user two programmed modes of dealing with them, (uncolored = entry on user’s list of scores or uncolored = “average of other colors in row”), or the option of stopping to color them manually.

10 If the user designates a column of individual compound labels as the “cluster labels,” then the system compares single compounds rather than clusters.

The mechanism operates as follows, with reference to FIGURES 11A-11C.

- (1) On the “Cluster Control” sheet 314 the user picks a control panel (e.g., panel 1100) to use (a panel which is empty or one containing non-
15 needed inputs). On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input information for several tools.
- (2) In the top black section 1102 of the selected control panel 1100, the user gives this new parameter set a name if not already done. The
20 name will be used to label the outputs.
- (3) In the blue section, the user enters a sheet name (in 1104) and the column (in 1106) to be considered as the cluster labels. Note: To score each compound separately rather than in clusters, enter a column with individual compound labels as the “Cluster Col.”

(4) The red section of the control panel is divided into two parts, with its action button **1108**, with red text “Score and Sort Clusters”, in the middle. Above the button, enter the names of the colors **1110** to be assigned point scores, along with their corresponding point scores **1112**. The scores are arbitrary and relative; they will be normalized by the system as necessary. However, a user should be sure always to assign higher point scores to colors which denote favorable values, and lower point scores to colors which denote unfavorable values. The cells with entries need not be colored, and need not be in score order, because the system will color and sort these cells when run.

When assigning point values, a user should be aware that uncolored cells (which are most likely blank, i.e., unknown data) may have quality values above or below those that contain grouped and colored data.. The user may decide that some of the colored groups are “better” or “worse” than data being unknown, and can assign a score to the color “NONE” accordingly.

(5) Below the “Score and Sort Cluster” button **1108**, the user enters the columns **1114** to use for scoring (using the same syntax as for the Data Coloring) and their corresponding relative weights **1116**. The numbers for weights are arbitrary and relative; they will be scaled by the system as necessary. Note that a line with multiple columns will assign the entered weight to *each* of the columns.

(6) The user the selects (clicks) the red-text “Score and Sort Clusters” button **1108**.

(7) When the scoring and sorting tool runs (on the system 100), if the system detects uncolored cells in the data, the user will be offered two modes of dealing with them automatically, or a third manual option of stopping to color them. The two modes are:

- 5 • “Use score for the color "none" on my list”
- (RECOMMENDED)
- “uncolored = average of other colors in row”.

(8) The program then scans the chosen columns in each row and adds up
10 the chosen column’s color scores for that row. These scores are then
averaged for each cluster of rows, as defined by the user-selected
“cluster column.” The VB program code which accomplishes this is as
follows:

```

15     For icol = 1 To ncols
         colorcode =
         Cells(irow,colnum(icol)).Interior.ColorIndex
         colorfound = False
         ' Add up the weighted scores
20     For j = 1 To ncolors
         If (icolor(j) = colorcode) Then
             jscore = j
             colorfound = True
             Exit For
25     End If
         Next j
         If colorcode = xlNone And treatblanks = "AVERAGED"
         Then colorfound = True
         If Not colorfound Then
30         Cells(irow, colnum(icol)).Select
         If colorcode = xlNone Then

```

```

        thiscname = "none"
    Else
        Call ColorNameToIndex(thiscname, colorcode, True)
    End If
5   addscore = score(jscore) * colweight(icol)
    cmpdscore = cmpdscore + addscore
    ' Next IF-THEN-ELSE block is
    ' special calculations for the "averaged" mode
    If colorcode = xlNone Then
10      lostweight = lostweight + colweight(icol)
        minscore = Application.Min(minscore, 0)
        maxscore = Application.Max(maxscore, 0)
    Else
        cmpdscore2 = cmpdscore2 + addscore
15      End If
Next icol

```

(9) The scores are then normalized for the various cluster sizes (number of rows per cluster), and scaled to a value of one hundred (100) for a row which is colored entirely with the user's highest-scoring color and a value of zero for a row that is colored entirely with any color to which the user has assigned a score of zero.

```

If clusterscore(icluster) = 0 Then
25   ' do nothing
ElseIf clusterscore(icluster) > 0 Then
    clusterscore(icluster) =
        100 * clusterscore(icluster) / (nrows * maxscore)
ElseIf clusterscore(icluster) < 0 Then
30   clusterscore(icluster) =
        100 * clusterscore(icluster) /
            (nrows * (-minscore))
End If

```

(10)The results are presented as two newly inserted worksheets. The first is named by appending the word "SCORES" to the name of the original data sheet, and contains a list of the clusters with their sizes and scores.

5 (11)The second new sheet is named by appending the word "SORTED" to the name of the original data sheet. The "SORTED" sheet contains a copy of all the original data and coloring, but with the rows re-ordered to place the highest-scoring clusters at the top, and all the clusters in descending score order from there down.

10 (12)The user has two additional options regarding the appearance of the "SORTED" sheet: (a) a column containing the numerical scores can be added; and (b) the columns that were not used in the scoring can be hidden, so that only the ones actually used remain visible.

15 An example of user provided data is shown in the control panel in **FIGURE 11D** which is taken from the cluster control worksheet shown in **FIGURE 11A**. As shown in **FIGURE 11A**, the parameters are stored with the name "Cmpd" **1102**. The scoring a sorting parameters in the control panel **1100** of **FIGURE 11D** give the color red a score of "-1", orange has a score of "0", yellow has a score of "1" and light green has a score of "2". Columns C and D have relative weights of "1", as does column E.

20

Note on the output of score and sort clusters: The system inserts two new sheets after the data (see, e.g., **FIGURES 14C-14D**). The first added sheet contains two score columns: the scores generated by *both* of the auto modes (uncolored = zero and uncolored = average), but the one not selected will be gray. The scores

25

are on a scale of “-100” to “+100”, where a score of “-100” means that all cells had the maximally negative score available, and a score of “+100” means that all cells had the maximally positive score available. The second added sheet has clusters sorted according to the *one* auto mode chosen when the tool ran. The routine offers to hide all columns that were *not* used in the scoring and sorting.
5 The user can selectively unhide certain columns by using the “Edit:GoTo” menu option (or typing “CTRL-G”), enter the columns in the “Reference” box (for example, C:F), then pick the “Format:Column:Unhide” menu option.

If the user wants to see a color-score-sorted list of compounds within a particular cluster (such as the best cluster), the user should do the following:
10
1. Sort by clusters to find the ID of the cluster wanted.
2. With a second rule, sort by compounds.
3. Go to the “SORTED by Compound” results sheet and turn on EXCEL’s “Data:Filter:AutoFilter” feature for the column that specified the clustering in the first sort. The user can then choose to view only the compounds in one particular cluster, and they will be in compound-sorted order.
15

Example

20 With reference to the already-colored worksheet “DEMO 1” shown in FIGURES 8A-8B, the cluster control worksheet shown in FIGURE 11A, and the control panel shown in FIGURE 11D, application of the scoring and sorting of clusters is described. As noted above, in the control panel of FIGURE 11D, the parameters are stored with the name “Cmpd” 1102. The color red has a score of

“-1”, orange has a score of “0”, yellow has a score of “1” and lightgreen has a score of “2”. Columns C and D have relative weights of “1”, as does column E.

Application of control panel “Cmpd” of **FIGURE 11D**, by selecting “Score and Sort Clusters”, produces the worksheets shown in **FIGURES 14A-14B**. When
5 the user selects the “Score and Sort Clusters” button **1108** for the “Cmpd” control panel of **FIGURE 11D**, the system first presents a dialog box (**1402** shown in **FIGURE 14A**) asking the user how un-colored cells should be scored for sorting. As noted above, un-colored cells can be scored explicitly by user entries (recommended) or as the average of the colors in the same row. Once the user
10 makes a selection and clicks on the “OK” button, the system scores and sorts the data, producing the display screen shown in **FIGURE 14B**. The system provides a summary of what was done, including the information about the two new sheets (“DEMO 1 SCORES by Cmpd” and “DEMO 1 SORTED by Cmpd un=ze”) which are added to the workbook. **FIGURES 14C-14D** show the data in the newly
15 created worksheet “DEMO 1 SCORES by Cmpd”.

Example

With reference to the already-colored worksheet “DEMO 2” shown in **FIGURES 10A-10B**, the cluster control worksheet shown in **FIGURE 11A**, and the control panel shown in **FIGURE 11C**, application of the scoring and sorting of
20 clusters is described. In the control panel of **FIGURE 11C**, the parameters are stored with the name “acids” (**1102**). The color red has a score of “0”, orange has a score of “1”, yellow has a score of “2” and light green has a score of “3”. Column D has a relative weight of “1”.

The application of the parameters or rules in the “acids” control panel produces two new worksheets (“DEMO 2 SORTED by acids” and “DEMO 2 SCORES by acids”) shown in FIGURES 15A-15B.

5 4. Score and Sort Dose-Response Data.

Data grouping and visualized by color coding has also been found to enable an automated solution to another vexing pattern recognition problem. An HTS lab is currently able to provide dose-response data on some subset of the whole collection of compounds originally tested. Sometimes, logistical
10 constraints (time and/or cost) dictate that only a few concentration points can be run on each compound, and the high-throughput nature of the process generates somewhat noisy data. A similar situation sometimes exists in other biological laboratories where assays are very time-consuming. Dose-response curves with few, noisy points are difficult to analyze by traditional curve-fitting methods. The
15 present invention includes a mechanisms/algorithms for analyzing percent-of-maximal-effect data and accurately ordering the compounds by potency, even when faced with few points and high noise.

The mechanism recognizes two properties of the dose-response data for each compound:

- 20 1. “Dose-responsiveness,” the drop-off of activity with dilution, is taken as a sign that the compound has some reasonable pharmacological mechanism of action.

2. The activity measurements at the various concentrations also provide a confirmation of the general level of each compound's activity that was indicated by the original single-poke HTS hit.

These two properties are somewhat independent, as illustrated by the example of a compound that is 95% active at all tested concentrations. It demonstrates very poor (i.e., no) dose-responsiveness over the range of concentrations tested, but is so active that it should not be ignored, because it might reveal a dose response if tested at even lower concentrations.

By using the data groupings and color codes of the dose-dependent activity data columns, which help to smooth out excessively fine distinctions in the numbers, this invention includes an algorithm to assign numerical scores for dose-responsiveness and overall activity in the dose-response data. Moreover, the algorithm also calculates a smart composite of these two scores, in such a way that a highly active compound will get a high composite score even if its dose-responsiveness is poor. This composite score is capable of extracting useful information, even from very noisy data, and has been validated to correctly order a list of test compounds. The system of this invention adds data columns that report all three scores for each compound, and these columns can themselves be color coded, and thus used in further comparison to other types of data by compound or cluster scoring and sorting as described above.

Moreover, within certain limits, the invention's dose-response scoring algorithm can also be used to make *quantitative* estimates of IC_{50} values of compounds, even in the presence of large amounts of experimental error. This is accomplished by adding a set of hypothetical marker compounds with known potencies and theoretically calculated activities at the test concentrations. Since

the ordering algorithm is reliable, these markers will be ordered into their appropriate place, and can be used to calibrate the ordering scores in terms of actual IC_{50} 's. In other words, estimates of IC_{50} for the compounds can be generated by interpolating between the markers in the ordered list of composite scores.

Scoring and sorting dose-response data according to the present invention processes several columns of colored dose-response data (activity vs. concentration) to assign three numerical scores that can later also be color coded, and thus used by the "Score and Sort Clusters" mechanism (described above) to compare compounds or clusters of compounds. The three scores are:

- (a) degree of dose-responsiveness over the concentration range tested;
- (b) overall activity level demonstrated in the dose-response data; and
- (c) a variably weighted composite of (a) and (b), designed to give high scores for high activity even when dose-responsiveness is poor (e.g., a compound that is highly active at all concentrations).

The scoring and sorting dose-response data according to the present invention bases its scoring on colors rather than absolute activity numbers. The mechanism takes user input from a *Cluster Control* worksheet, e.g., as shown in **FIGURES 11A-11B**. **FIGURE 11B** shows a control panel from the cluster control worksheet shown in **FIGURE 11A**, wherein the user has selected columns F to I of worksheet "DEMO 1" for scoring dose-response.

The system detects uncolored data, notifies the user, and asks whether to continue. If yes, the system skips the row containing the uncolored data. The system inserts three new columns on the original spreadsheet to contain the new scores, the new columns immediately following the columns of dose-response data. The column headings show the name of the parameter set used for scoring. Preferably, the system offers to regenerate existing table of Sheet Statistics to correct it for newly added score columns. Further, the system offers to sort the data rows by decreasing score. The system also offers to carry out quantitative estimates of IC_{50} values for the user's compounds, by adding artificial calculated calibration marker compounds.

In order to score and sort dose-response data:

- (1) Ensure that the dose-dependent activity data columns are ordered with highest concentration at the left and lowest concentration at the right. To ensure this, the system will remind the user of this requirement and ask him to confirm it when this tool is run. *Note:* If the data are for an undesired effect such as toxicity, the columns should be ordered the opposite way (lowest concentration left, highest right).
- (2) Use the Data Coloring (described above) to color the dose-response data columns.
- (3) Go to the Cluster Control sheet 314 (FIGURES 11A-11D) and pick a control panel 1100 to use. On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input information for several tools.

- (4) In the top black section, give this new parameter set a name (1102) if not already done. The name will be used to label the outputs.
- (5) In the blue section, enter a sheet name (1104).
- (6) In the red section (1110), enter the colors used to color the dose-dependent data, and relative point scores (1112) to be assigned to these colors.
- (7) In the green section (1118), enter the columns which contain the dose-response data (using the same syntax as for Data Coloring).
- (8) Click the green-text "Score Dose-Response" button (1120).
- (9) If the data are expressed as "percent of maximal effect," the user can follow the prompts to add calibration markers and make quantitative estimates of IC_{50} 's.

Note on the output of score and sort dose-response: the system inserts three score columns after the dose-dependent data. The three scores are all scaled to a 0-100 range, and have meanings as follows:

- (a) degree of dose-responsiveness over the concentration range tested:
- 100 = smoothly decreasing with dilution, spanning the entire range of color groups;
- 75 = flat dose-response; and
- <75 = even more poorly behaved
- (b) overall activity level demonstrated in the dose-response data
- 100 = highest activity color group at all concentrations.

(c) a variably weighted composite of (a) and (b), designed to give high scores for high activity even when dose-responsiveness is poor (e.g. a compound that is highly active at all concentrations).

5

The Dose-Responsiveness Scoring Algorithm

The data columns are ordered left to right, by decreasing concentration.

The scoring algorithm awards positive score points for each dilution step across the data that actually shows a decrease in the activity data *group* (i.e., the color),

10 and to penalize every step that does not. The algorithm uses the following scoring:

- +1 point (awarded) when a dilution step moves to a lower activity group
- 0 points when a dilution step leaves the activity group unchanged
- -3 points (penalty) when a dilution step moves to a higher activity group

15

The maximum score would go to a compound that shows all the possible color group steps in the right direction, and has no reversals. The minimum score would go to a compound with all the possible reversals, and no correct steps. The program then scales the extremes to 100 and 0, in order to present a consistent interface to the user.

20

The relative magnitudes of the scoring parameters were empirically arrived at by testing "complete sets" of color patterns. This is possible because of the data simplification afforded by the value grouping. If we define the following numerical parameters:

$C \equiv$ number of colors used, i.e., number of data value groups

$P \equiv$ number of points measured, i.e., number of different concentrations
(doses) tested,

then the entire “universe” of possible color patterns includes (C^P) different cases.

For some typical values that might be encountered in real HTS data, this total

5 number of cases is manageable in EXCEL, as shown by TABLE 1, below.

TABLE 1. Total Number of Color Patterns

P = # of conc. Points	C = # of color groups	total number of possible cases
3	3	27
3	4	64
4	3	81
4	4	256
5	3	243
5	4	1024
6	3	729
6	4	4096
7	3	2187
7	4	16384*

* For a spreadsheet with a heading row, this exceeds EXCEL's current limit (for EXCEL 95) by one. This value should not exceed the limit for Excel 97.

10 Scoring was done on several of these complete sets. In each set, the results were sorted by decreasing score and compared to “intuition” for general correctness of ordering of dose-responsiveness, and scanned for cases deemed to be clearly out of order. The (+1, -3) score set was found to produce satisfying ordering, while lesser penalties led to poorly ordered results. More objective tests
15 of ordering (described below) were then used to further validate the algorithm

The case of $P=3$ and $C=3$ is presented below in its entirety for illustration.

TABLE 2 (FIGURE 16F) shows artificial data and processing for twenty seven (27) hypothetical compounds. The “percent inhibition” columns represent assay “data.” If one defines three groups by breakpoints at 33% and 66%, each cell is

assigned to a data group as shown in the middle set of three columns. Here it is clear that the order of compounds in this table is systematic (111, 112, etc.), to illustrate that the complete set is present. The third set of three columns shows color coding, with the darkest being least active and the lightest being most active.

5 Then the data set was processed by the system to yield dose-responsiveness scores, and the results sorted by this score, giving **TABLE 3 (FIGURE 16G)**, the complete set in order of decreasing dose-responsiveness. **TABLE 3** also shows the intermediate step-scoring and unscaled score points, to aid in following and understanding the algorithm. These points are not displayed
10 by the system itself.

The Overall-Activity Scoring Algorithm

The second property of interest to be extracted from the data is the overall activity level exhibited by each compound. As explained above, this is largely
15 independent of the dose-responsiveness.

The data value groups' ordinal index numbers are used as single-point activity measures instead of the original data numbers. Extra weight is given to activity shown at lower concentrations by the simple algorithm of weighting each data column by its serial position, again ignoring the actual concentration values.
20 The scores are then scaled to the range 0 to 100. The results of this scoring on the same complete set are shown in **TABLE 4 (FIGURE 16H)** which has been re-sorted by decreasing overall activity.

The Composite Scoring Algorithm

Comparison of Tables 3 and 4 (FIGURES 16G & 16H) shows clearly that the compound ordering by dose-responsiveness is quite different from the ordering by overall activity. The user (a chemist) could now color-code the new score columns and use them as independent factors in a larger scoring. However, chemists also want a *single* index of compound quality derived from the dose-dependent data. Moreover, a composite index would further help to alleviate the effects of noise on data interpretation, by incorporating more information into the ordering process. This is an “information-based smoothing” of the data. Therefore, a procedure to calculate a third, “smart composite” score from the other two scores was devised.

The general idea is that when selecting good compounds from dose-response data, compounds showing overall high activity should not be discarded for lack of responsiveness. Therefore, the smart composite score should give more weight to the overall activity when the overall activity is high, but lower weight when it is low. A generalized weighted average is written as

$$\text{composite score} = (\text{activity weight})(\text{activity score}) + (\text{responsiveness weight})(\text{responsiveness score})$$

or, defining corresponding symbols:

$$S_C = (W_A)(S_A) + (W_R)(S_R)$$

If the weights are normalized to sum to unity, then this becomes

$$S_C = (W_A)(S_A) + (1 - W_A)(S_R)$$

The activity weight W_A varies with the activity score S_A in such a way as to achieve the desired result.

The functional form of this variation was the subject of empirical testing. It was decided that the limits would be that W_A would approach 0.5 (activity and responsiveness equally weighted) in the limit of low activity, and that W_A would approach 1.0 (responsiveness ignored) in the limit of high activity. The actual
5 variation was encoded as an exponential increase in order to have rather sharp onset of the activity bias at higher activities:

$$W_A = (C_1) \exp [(k)(S_A)] + C_2$$

The value of the coefficient $k=0.06$, for which the activity bias starts to become substantial around an activity score of eighty (80), was chosen for
10 implementation in a preferred embodiment of this invention, according to empirical results. FIGURE 16I shows the variation for a few values of k . TABLE 5 (FIGURE 16J) shows all three scores for the example complete set, now sorted by decreasing composite score.

15 The details of the scoring algorithms were arrived at largely by comparing results to intuitive ordering of the test cases in the complete sets. Because the sets were complete, no really new results can be generated by further test sets. However, one can generate test activity data sets from compounds of known potencies, whose real rank ordering is thus known, in order to see more
20 objectively how well the scoring algorithms rank the results.

To this end, a set of pseudo-ligands was hypothesized, with dissociation constants from a fictitious receptor ranging from nanomolar to millimolar ($pK = 9$ to 3). The set included thirty one (31) compounds, with potencies evenly spaced by 0.2 log units (9.0, 8.8, 8.6, ... , 3.4, 3.2, 3.0). A "pseudo-screen" was created
25 which "tested" binding of these ligands at five concentration points in the usual

range: 10^{-5} M, 3×10^{-6} M, 10^{-6} M, 3×10^{-7} M, and 10^{-7} M. Note that the span of potencies exceeds the span of concentrations tested by two log units on each end, so the test set includes both “very active” and “very inactive” compounds relative to the screening concentrations.

Then artificial binding data were created by calculation as follows.

Assuming a simple binding equilibrium of the ligand to a receptor, the “percent inhibition” at a given ligand concentration is equal to the fraction of receptor sites which are occupied by ligand, given by simple equilibrium equations as

$$10 \quad p_{\text{inhib}} = 100 \bullet (\text{ligand}) / [\text{K} + (\text{ligand})]$$

For a more realistic simulation, artificial random noise was then added to the calculated numbers. The first experiment reported below used noise randomly distributed over the range of ± 10 inhibition percentage points, and the second with noise up to ± 30 inhibition percentage points. Note that this means ten or thirty percentage points of absolute error, not 10% or 30% of the value.

The artificial data were then color-coded according to the mechanisms of this invention (described above) into four color groups, using the simple breakpoints at 25, 50 and 75 percent inhibition. Note that in assigning these breakpoints, no consideration was given to the actual data values. Then the scoring algorithms of this invention were run, and the compounds sorted by the composite score. Rank order numbers were assigned to the compounds, with 1 being the most potent and 31 the least. In cases of ties in the composite score, equal rank numbers were assigned, with a value equal to the average of the rank numbers spanned by the tied group of compounds (e.g., a tie for 2nd and 3rd

would result in each compound being ranked "2.5"). For each experiment, the final rankings were plotted against the "real" rankings by known potency, to test how well the scoring algorithms ordered the compounds. These plots are shown in **FIGURE 16K** (for noise=10) and **FIGURE 16L** (for noise=30).

5 For the experiment with noise up to 10 inhibition percentage points, shown in **FIGURE 16K**, the ranking of the composite scores is "perfect" (in the sense of having no inversions) over the range of tested concentrations ($pK = 5$ to 7). The pseudo-screen is unable to distinguish the potencies of compounds above or below this range.

10 When the noise is much higher (30 percentage points), the ranking of individual compounds is not as precise, but one can identify three cleanly divided "good-medium-bad" groups, as indicated by the dashed boxes on **FIGURE 16L**. Thus, even with this rather extreme noise level, the invention's scoring still successfully prioritizes the compounds into groups. The range where
15 discrimination is effective is still roughly the range of the test concentrations ($pK = 5$ to 7), but has been reduced somewhat by the higher noise. Note that the ranking within this range (the middle boxed group) is still mostly correct, with only one inversion, even for single compounds.

Quantitative Estimation of Potencies

20 With confidence established that the algorithms provide reliable rankings of compounds by potency, it is possible to proceed to making quantitative estimates. The method uses calibration marker compounds.

To understand this method, it is helpful to realize that the concept is analogous to the quantitative use of SDS polyacrylamide electrophoresis gels to

measure protein molecular weights. The proteins are known to migrate through the gel with speeds directly dependent on molecular weight, but it is difficult to calculate the absolute migration rates for a particular experiment. In dose-response scoring, the compounds are known to be properly ordered, but it is not clear how to calculate a potency (e.g., K_{diss} or IC_{50}) directly from the score.

Protein chemists solve the molecular weight problem by running marker proteins, with known molecular weights, in the same gel, then using their band positions as calibration for the unknowns. Analogously, this invention's quantitative estimation method uses hypothetical marker compounds of known potency to internally calibrate the dose-response composite scores for the user's choice of a coloring rule, then interpolates the potencies of the unknowns.

To create markers, the system asks the user to input the concentrations used for each of the dose-dependent activity data columns. The system then picks a set of calibration concentrations, at intervals of 0.5 log units, to span the tested range. For each of these calibration concentrations, a marker compound is created and added to the user's compound list, and artificial data is calculated for each column, from the same simple equilibrium binding equation used above in the validation study (this time with no "noise"):

$$p_{inhib} = 100 \bullet (ligand) / [K + (ligand)]$$

The marker data are then colored by the same rule used for the user's compounds, and the scoring and sorting algorithm is re-run.

The result is that the markers are sorted into the list according to their potencies, and the potencies of the other compounds can be estimated by

interpolating between the markers, using the composite dose-response scores. To illustrate, a typical section of a sorted list is shown below in **TABLE 6** (**FIGURE 16M**), using four colors.

Potencies for compounds that fall between two markers are calculated by linear interpolation between the logarithms of the markers. Given the various uncertainties in the data values themselves and in the evaluation process, it was found that linear interpolation between markers spaced at 0.5 log unit intervals was sufficiently precise, and no more complex curve fitting was necessary.

Validation of Quantitative Estimation

Validation of the quantitative estimation method followed a procedure very similar to that used to validate the scoring, and using the same sets of test data with various noise levels. As before, the testing concentrations were from 10^{-5} to 10^{-7} M (negative log from 5 to 7). Marker compounds (no noise) were added with pK's from 4.5 to 7.5, and K_{diss} estimates for the noisy compounds were carried out by the interpolation method. The results are shown below for the cases of 10 and 30 inhibition percentage points of noise.

FIGURES 16N and 16P show that the estimates are clearly quite good within the range of the testing concentrations (pK 5 to 7), but the quality of estimation deteriorates quickly beyond those limits, and algorithm does not reliably distinguish among compounds whose potencies are more than a half log unit beyond the testing range. Therefore, it was decided that presently preferred embodiments would not report any estimated values that fell outside the range of concentrations used in the testing data columns. Thus, in the example in **TABLE 6** (**FIGURE 16M**), the lowest testing concentration was 10^{-7} M (= 0.1 μ M). For the

first compound in TABLE 6, the system has estimated a potency with $pIC_{50} > 7$, but it conservatively only reports " $<0.1 \mu M$."

TABLE 7 summarizes the statistics of the estimations within the testing limits. TABLE 7 shows that the method successfully estimates the potencies within about a factor of two, even with high noise levels.

TABLE 7. STATISTICS OF ESTIMATION VALIDATIONS

<i>percentage points of noise</i>	<i>number of compounds</i>	<i>average of abs(log error)</i>	<i>Average error in IC_{50} (factor)</i>
10	13	0.22	x 1.6
30	13	0.39	x 2.5

Comparison to Other Methods of Quantitative Estimation

Further corroboration was obtained by treating some real data from T-cell proliferation blockage assays. It is estimated that these data have *at least* as much noise as the artificial test set with 30 inhibition percentage points added. The standard treatment of this data in the past has been to fit a dose-response curve with a Hill coefficient of 1, using a PC-based program ORIGIN. (ORIGIN is a data analysis program from Microcal Software, Inc. of Northampton, Massachusetts. ORIGIN is used in this instance for non-linear least-squares fitting of dose-response curves to functional equations.)

The data used here were from testing in the concentration range from 1 to 0.03 μM (negative log from 6 to 7.5). The plot in FIGURE 16Q shows the correlation of values estimated by this invention with values from ORIGIN fits.

Two compounds that the present invention estimates to be beyond the testing range, i.e., pIC_{50} below 6, are included as open diamonds, for illustrative purposes explained below. (As explained above, preferred embodiments of this invention would normally not report these values.)

5 The results are consistent with the properties observed in the validation study. The present invention estimates are quite good within the testing range (6 to 7.5). At the lower limit, this invention has made two estimations exactly at 10^{-6} M (arrows) which do not correlate as well with the ORIGIN fits. Nevertheless, because the whole plot spans only a relatively narrow range of potency, even these
10 discrepancies are not very large. For all twenty eight (28) estimates within the testing range (including these two), the average logarithmic deviation between the two estimation methods is 0.18, corresponding to a factor of only 1.5.

 It is further noted that the calibration marker estimation method does not uniformly “flatten out” beyond the testing concentration range. The two open
15 diamonds in **FIGURE 16Q** are estimations that presently preferred embodiments of this invention would not normally report because they have $pIC_{50} < 6$, but they agree well with ORIGIN fits

 It is interesting to compare calibration-marker with curve-fitting results for particularly badly behaved data, such as dose-response curves that are not
20 monotonic with respect to concentration. This is sometimes the type of data that emerges from dose-dependent screening in a high-throughput mode. Studies of this type have been initiated by adding artificial noise to the extent of fifty (50) inhibition percentage points.

 Finally, it should be pointed out that there is some mechanical advantage
25 to using the present invention relative to current practice of using ORIGIN.

ORIGIN is used by manually filling in a template with data, then manually
executing a fit. Depending on the number of points and the degree of
customization of parameters, this can take one to ten minutes of the user's time.
The present invention, on the other hand, processes a whole spreadsheet at once
5 (i.e., up to 16,383 compounds), and goes at a rate of about 3,000-4,000
compounds per minute on a 200 MHz PC.

Examples

FIGURE 16A shows dose response data for twenty (20) compounds at four
10 concentrations. The data have been grouped and the cells colored by the rule
shown in FIGURE 16B. The result of the scoring and sorting process is shown in
FIGURE 16C, where the compounds are ordered by decreasing values of the
composite score (column H). Then, virtual "marker" compounds are added with
known potencies spaced by 0.5 log units, and they are shown in FIGURE 16D,
15 colored by the same rule and scored. The name of each marker compound
designates the logarithm of its potency, e.g., "marker_7.0" has a potency $IC_{50} = 10^{-7}$ M. FIGURE 16E shows the result of sorting the list by decreasing composite
score after adding the markers. This process then enables estimation of IC_{50}
values for the compounds by interpolating in the column (H) of ordered composite
20 scores, and these estimates appear in two forms in columns I and J.

6. Summarize Spreadsheet Statistics Mechanism

This mechanism creates a table summarizing the entries in each column of
a data sheet, to aid the user in deciding how to color each column. The

mechanism counts numeric, text, and data entries, and uses color to flag columns that have mixed types. The mechanism also counts blanks, and specially flags columns with "trailing blanks," i.e., columns which are shorter than the longest one on the spreadsheet. For numeric data, the mechanism calculates minimum, maximum, mean, and standard deviation, even in the presence of interspersed text entries. For text data, the mechanism presents a list of the text strings used and their occurrence counts. The mechanism creates a summary key of the column letters and headings as a text box that can be copied to other sheets for convenient reference.

FIGURE 17A shows a sample spreadsheet containing miscellaneous data on twenty four (24) compounds. **FIGURE 17B** is the statistics sheet calculated from it. Each row of the statistics sheet describes one column of the original data sheet. First, the counts of numeric, text, date, and blank entries are listed, followed by two columns describing the total length of the data sheet. Then the minimum, maximum, mean, and standard deviation of any numeric data are reported. Finally, the statistics sheet lists a summary of the text strings found in each original data column. As examples, in **FIGURE 17B**, one can see that original column A ("Cmpd") had twenty four (24) different text strings, that the numeric data in original column C ("Test1") had a mean of 2.385, and, flagged by the red coloring, that original data column E ("Test3") had a mixture of ten numeric data and two text strings, both "N.A."

The details of how the program code accomplishes this are straightforward, and one of ordinary skill in the art would know, from this description, including the Figures, how to make and use this invention. The

program loops through all the entries in the column, testing the data type of each, and tallying the counts and numerical statistics.

Spreadsheet Creation and Organization

5 The operations of this invention require a considerable amount of user input, e.g., to create well-structured spreadsheets, to define and apply diverse coloring rules for large numbers of columns, and to use these colors and the user's stated scientific priorities to create meaningfully ordered lists of compounds or clusters.

10 The user interface of this invention has been designed to ease this process and help the scientist focus on the tasks of formulating and recording clear descriptions of the evaluation parameters. Accordingly, this invention provides a number of tools and mechanisms to aid in the creation and organization of spreadsheets. These tools and mechanisms include:

- 15 • Smart Append Column Mechanism
- Merge Data Mechanism
- Data Import Mechanism
- Workbook Navigation Shortcuts
- Conversion of "uM" to μ M and "UU" to μ
- 20 • Delete Pictures Mechanism
- Change Values in Column Mechanism
- Concatenate Values across Columns Mechanism
- Delete Leading Inequality Signs Mechanism
- Delete Derived Sheets Mechanism

Smart Append Column

This mechanism appends new columns of data onto an existing spreadsheet, matching rows by labels (e.g., compound numbers). The mechanism
5 copies all data to a new sheet before doing its work, leaving the original sheets unchanged. There is no need for the user to pre-sort any of the data. The mechanism provides optional case-sensitive or case-insensitive label matching.

New rows are added at the bottom when new labels do not match any old labels. Rows with missing labels are identified and the system offers to fill them
10 by copying previous label. Rows with repeated labels (i.e., replicate data) are also identified and the system offers a choice from among several automated processing rules, or manual fixing. A fast matching algorithm temporarily sorts rows by label, then restores original order when finished. Several intermediate stopping points are offered and extra data viewing options for conservative users
15 worried about errors.

Merge Data Mechanism

The merge data mechanism copies new data values from an appended
20 column into an older column. The mechanism copies all data to a new sheet before doing its work, leaving the original sheets unchanged. The mechanism detects cells where new data would overwrite old data that is different, flags them with color, and alerts the user. Several intermediate stopping points are offered to

the user, as are extra data viewing options are offered for conservative users worried about errors.

Data Import

5 One-button (or one-menu-click) import of existing EXCEL spreadsheets into an integrated file, which contains both the data and the related control sheets. The mechanism offers to search for and remove any leading or trailing spaces in the imported data and offers to consolidate replicate data rows into unique ones, using user choices as to how to handle the replicate data. The mechanism also
10 detects hidden rows and offers to unhide them and detects formulas and offers to convert them to values. This mechanism is also used to update to newer version of the system.

Workbook Navigation Shortcuts

15 The system includes various workbook navigation shortcuts including:

- A special added drop-down menu which includes commands for jumping directly to the various control sheets. These commands also have keyboard shortcuts assigned to them.
- From a cell on a control sheet that contains the name of a data sheet, a
20 special item on the right-click shortcut menu jumps directly to that data sheet. Other special items on this menu enable a “Twin Screen” display to see two sheets at once.

- To aid in choosing columns to enter on control sheets, there is a special “Twin Screen” display triggered by right-clicking any cell on a control sheet that contains the name of a data sheet.

5 Convert “uM” to μ M and “UU” to μ

Preferred embodiments of the system of this invention require the data spreadsheet to have *one and only one* row of column headings. The user can type either of the encoded strings “uM” (lowercase u, uppercase M) or “UU” (both uppercase) into any column heading, select the cell or whole row of headings, then
10 pick this command. Each “uM” in the selection will be converted to “ μ M”, and each “UU” will be converted to a “ μ ”. The code recognizes the special exception of the word “VACUUM” as long as it doesn’t end with the cases “uM.” This conversion allows the user to avoid the confusing use of lowercase “u” or the column-widening use of the full prefix “micro.” This utility appears on the
15 system menu.

Delete Pictures

The system provides a mechanism for removing pictures containing chemical structures, in order to reduce file size, processing time, and confusion
20 when they do not align properly after row sorting.

Change Values in Column

This is a mechanism for regularizing data in a spreadsheet column. It facilitates replacement of all occurrences of a given value by another. The mechanism creates backup copies of the original column, and updates any existing data statistics for the edited sheet.

Concatenate Values across Columns

The system provides a mechanism for regularizing data in a spreadsheet column. Some possible uses include: (a) construction of unique row labels: M-number plus stroke number → "M123456/001"; and (b) reconstitution of numerical inequalities from separate columns: ">" plus a number → ">number". The user is provided with an option to include linking (delimiting) text strings between values and an option to include or skip blanks. The system retains the original columns and inserts a new one for the results.

Delete Leading Inequality Signs

Another mechanism for regularizing data in a spreadsheet column includes the mechanism to delete leading inequality signs. This mechanism converts entries like ">1000" to just the number "1000". This must be used with considerable caution, because it is the equivalent of creating a false test result. It is generally preferable to color the cells containing text strings with the data coloring mechanism described above, rather than alter them. All later processing is based on the colors, not the cell values. This mechanism also deletes inequality

sign only if it is the first character in the cell. The mechanism creates backup copies of the original columns.

Delete Derived Sheets

5 The system menu includes a command to delete all output sheets from the current workbook, with separate user confirmation for each one. This is intended as a cleanup mechanism for information that may be outdated and is easily regenerated by subsequent system runs.

10 Initial experience with the coloring tool has revealed that color coding has more subtle, but far-reaching usefulness. The colors themselves also can act as a “currency of exchange,” a medium for comparing the quality of one kind of result to the quality of a very different kind of result. For example, an HTS activity of “95% inhibition” may be considered desirable and color coded, e.g., green. In the
15 same list of compounds, a molecular weight between 400 and 600 may be considered optimally desirable, and thus also color-coded green. If the user takes care when assigning colors, “green” takes on a common meaning across the board. This translation of data values into colors then opens up a cornucopia of possibilities for processing the colors (as numerical color indices) and comparing
20 compounds, searching, in our example, for the ones that are the “most green.”

 Accordingly, the system includes tools to numerically score individual compounds or clusters of compounds by the colors that appear in their various data columns. The system can then create a new spreadsheet sorted by this score

(either by single compounds, or cluster-by-cluster, the choice being the user's), in which the "most green" compounds will then appear at the top.

5 **Examples**

Application to Portfolio Management

The system can equally well be applied to any set of data where the rows are cases of a similar construct, with the columns being various properties of each case. For example, a data spreadsheet can contain a list of competing projects or
10 investments for a company's portfolio, with the columns containing various managers' ratings of each project or investment. **FIGURE 18A** shows an example of twenty projects, each of which has been scored 1, 2, or 3 on two factors, one more important than the other, by each of three managers. The sheet has been colored by the rule shown in **FIGURE 18B**. Then the data were scored and sorted
15 by the sorting rule of **FIGURE 18C**, and the result is shown in **FIGURE 18D**. Clearly, the projects that were given a "3" in the important factor come to the top, and it can be seen that the less important factor does indeed matter less to the final ordering. The colors also help to flag anomalies, such as a low score by one manager on an otherwise high-ranking project.

20 In general, the data can be various sorts of data. Some examples are listed below and illustrated in the referenced Figures.

FIGURE 19 shows a list of drug candidate compounds, scored and sorted by a composite of ten parameters that describe their physical, chemical, and biological properties. Green shades indicated desirable values; red shades are

undesirable. The display is compressed vertically with the vertical re-scaling tool to clearly display the difference in coloring patterns between the top eighty (80) compounds and the bottom eighty (80) compounds (separated in the illustration by a blank band).

5 **FIGURE 20** shows a list of proteins that are candidates for targets for drugs, chosen from a pool of candidate genes, scored and sorted by a composite of eleven parameters that describe their suitability.

FIGURE 21 shows a list of research projects competing for resources. Each project has been scored according to several evaluation factors, and the whole
10 array has been sorted by color groups. The same construct is useful for evaluating employee performance or job candidates.

FIGURE 22 shows a list of pharmaceutical companies and their current status with regard to discovering or marketing products in each of various disease areas. Each company's line of products has been scored according to the maturity
15 of the offerings, and the whole array has been sorted by color groups.

FIGURE 23 shows the use of data-grouping (coloring) rules to visualize the time courses of drug concentrations in blood. In this example, light colors were chosen to represent high concentrations of drug in the blood, while dark colors were chosen to represent low concentrations. The figure shows a wide range of
20 differing time courses.

FIGURE 24 shows the use of data-grouping (coloring) rules to visualize the matrix of pairwise cross-correlations of the results of eight (80) drug screens. In this example, light colors were chosen to represent low correlations, while dark colors represent high correlations.

Quantitation of the Similarity of Data Grouping in Two Variables

As part of the present invention, a mechanism is provided for assigning a quantitative measure to the degree of similarity of grouping (visualized by color coding) of data in each of two columns of an EXCEL spreadsheet. The mechanism
5 allows for a correlation-like analysis on a wide variety of data types, including text, or mixed numbers and text.

In the data-exploration paradigm of the present invention, one of the first steps a user takes is to divide the range of data values in each column into a small number of groups for further analysis, thus effecting a reduction of precision
10 which has been found to be useful in various ways.

It is sometimes useful to explore whether the rows of the data matrix have been divided into similar groupings in each of two different columns. For example, a researcher might ask, "Do the high molecular weight compounds tend to be the ones whose solubilities fall below the limits of measurement?" In other
15 words, this would mean to compare the groupings in the molecular weight column with the groupings in the solubility column.

If the data were strictly quantitative, this would be called *correlation* of variables, and there exist a number of perfectly good statistical measures of the phenomenon. However, one of the unique capabilities of the present invention
20 lies in dealing with textual data and mixtures of numbers and text, and it would be helpful if one could translate the visible color patterns of the present invention to some kind of quantitative measure of correlation. In order to avoid confusion with standard statistical correlation, the distinct term "color grouping similarity" is used to describe the new measure.

In preferred embodiments of this invention, the data grouping is stored in the form of the colored backgrounds of data cells. At first glance, one might simply seek to compare the colors of the first column with those of the second, and count the number of rows with matching colors. However, a color grouping
5 similarity tool must be able to cope with the possibility that the colors are different. This could happen because the user chooses completely *different color schemes* for the two columns, or because the correlation is negative. As an example of negative correlation, suppose column A contains random numbers between 0 and 1, colored such that those above 0.5 are green and those below 0.5
10 are red. Then imagine a column B where each value is equal to one minus the corresponding value in column A, i.e., the “one’s complement.” If the user colors the second column with exactly the same coloring rule as the first, every row will have a different color in column B than in column A. None of the colors will match, though the groupings are perfectly correlated. To be successful, the tool
15 must deliver a high measure of correlation between such pairs of columns. The algorithm described below was designed to perform in this way.

Algorithm for Measuring Data Grouping Similarity

The algorithm was derived from semi-quantitative reasoning, as follows.
20 It is based on the qualitative question, “For all rows that have one particular color in the first column, to what degree do they have a *uniform* color in the second column (not necessarily the *same* color as in the first column)?” The quantified answer to this question is then averaged over the set of colors used.

The details of the mechanism can be seen by example. First, to compare the grouping in two columns A and B, a matrix of “ordered color pair counts” (OCPC) is defined such that each matrix element $OCPC_{ij}$ is the count of spreadsheet rows where one finds color i in spreadsheet column A and color j in spreadsheet column B. Then, the rows of the two spreadsheet columns are scanned to count the number of occurrences of each ordered color pair and thus to determine the values of the matrix elements.

In the discussion below, carefully distinguish the rows and columns of the user’s *spreadsheet* of data from the rows and columns of the derived OCPC *matrix*.

If the two spreadsheet columns had *exactly the same coloring*, the nonzero elements of the OCPC matrix would all be on the diagonal. As a simplified example, consider four colors (green, yellow, orange, red) and a total of 16 data rows. The diagonal matrix might be (zero elements left blank for emphasis)

color in column A	color in column B →	green	yellow	orange	red
green		4			
yellow			2		
orange				7	
red					3

In the case above, there are groups of 4 spreadsheet rows colored green (in both columns), 2 rows colored yellow, 7 rows colored orange, and 3 rows colored red.

In contrast, if the groupings were the same, but the coloring rule for the second spreadsheet column used the same colors in a different order, the OCPC matrix might look like the following, no longer diagonal:

color in column A	color in column B →	green	yellow	orange	red
green				4	
yellow		2			
orange					7
red			3		

A simple extension applies if *entirely different colors* (cyan, blue, maroon, purple) are used in the second spreadsheet column. The OCPC matrix might then be:

color in column A	color in column B →	cyan	blue	maroon	purple
green				4	
yellow		2			
orange					7
red			3		

In *any* of the three cases above, the groupings are identical, and the OCPC matrices have the property that each matrix row and matrix column has only one nonzero element. That lone element is necessarily equal to the sum of the row or column. This situation should receive the *highest* similarity score.

One way to define the contrasting situation that would deserve the *lowest* similarity score would be that for each user-defined group of spreadsheet rows in one spreadsheet column, the other spreadsheet column has a “maximally non-unique” set of colors. In the corresponding OCPC matrix, this corresponds to each matrix row or column having a broad distribution of values rather than a single non-zero, a uniform distribution has been chosen as the definition of this state:

color in column A	color in column B →	cyan	blue	maroon	purple
green		1	1	1	1
yellow		1	1	1	1
orange		1	1	1	1
red		1	1	1	1

This low-similarity state can be more precisely defined by saying that each element in a given matrix row or column is the average of all the counts in that matrix row or column.

5 With these concepts defined and the OCPC matrix filled, the scores can then be derived. Each OCPC matrix row (corresponding to a color group in the first spreadsheet column) is selected in turn for scoring. Each element in the matrix row is given a score between zero and one, according to its linear interpolation between: on the one extreme, the average of the nonzero elements in
10 the row, and on the other, the sum of the row or column (i.e., the maximum value it could have if all the others were zero). The scores are then averaged over all the rows of the OCPC matrix to generate a row-wise score component.

Next, the corresponding process is applied to the *columns* of the OCPC matrix (each corresponding to a color group in the second spreadsheet column
15 rather than the first). The resulting column-wise score component is averaged with the row-wise score component, then the average is scaled to a maximum of 100 to generate the final similarity score for the two spreadsheet columns.

Interpretation of the Similarity Scores

20 Although the scores are quantitative and well-defined, their interpretation is best done in a partly subjective manner, based on experience. The behavior of

the scores is best understood by example. In **FIGURE 25**, the leftmost (“base”) column has been compared to each of the others, and the scores are shown as well as pictures of the grouping patterns. Comparison of the base with the next column shows that the tool delivers a maximal score of 100 for identical grouping, even when the colors are completely different. Then, stepping across the figure toward the right, it can be seen how the score decreases as the grouping pattern gradually becomes less similar to that of the base column. All the way down to a similarity score of 40, it is still basically true that the light colors are on top and the dark on the bottom, with increasing “noise,” but when the score falls to 20, the pattern appears to have no correspondence to that of the base.

Implementation of the Data Grouping Similarity Tool

In practice, the tool allows the user to choose two sets of spreadsheet data columns. The program then automatically generates all pairs containing a column from the first set with a column from the second set, then writes the similarity scores onto a newly inserted spreadsheet in the user’s workbook. The output takes the form of a table where the degree of similarity is itself color-coded to aid the user in identifying significant cases. An example appears **FIGURE 26**.

While the invention has been described with reference to particular mechanisms (algorithms, processes and functions) and architectures, one skilled in

the art would realize that other mechanisms and/or architectures could be used while still achieving the invention.

While embodiments of the present invention have been described with particular setup and initialization procedures, other setup and/or initialization
5 procedures can be used.

Further, while many of the operations have been shown as being performed in a particular order, one skilled in the art would realize that other orders, including some parallelization of operations, are possible and are considered to be within the scope of the invention.

10 While the present invention has been described with reference to analysis and pattern recognition in data sets relating to chemical compounds, the methods, systems and devices of this invention are considered to be general constructs covering other, non-chemical data sets.

Thus, are provided methods, systems and devices for analysis and pattern
15 recognition in large, multidimensional data sets using low-resolution data grouping. One skilled in the art will appreciate that the present invention can be practiced by other than the described embodiments, which are presented for purposes of illustration and not limitation, and the present invention is limited only by the claims that follow.

What is claimed is:

1. A method of operating on data, the method comprising:
providing at least one user-defined grouping rule for grouping the data into
5 a user-definable number of groups; and
applying at least one of the grouping rules to the data.
2. A method as in claim 1 wherein the data are provided in a table and
wherein the at least one grouping rule applies to at least one user-selectable
10 column of the table.
3. A method as in claim 1 wherein the at least one grouping rule
defines breakpoints corresponding to the user-definable number of groups, and
wherein application of the at least one rule to the data divides the data into groups
15 based on the breakpoints.
4. A method as in claim 1 further comprising:
presenting the grouped data in a manner that visually distinguishes the
groups.
20
5. A method as in claim 4 wherein the grouping rules associate colors
with groups and wherein the presenting of the grouped data further comprises:
coloring an aspect of the data according to the rules.

6. A method as in claim 4, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein the presenting of the grouped data comprises:

5 coloring each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

7. A method as in any one of claims 3 and 6, wherein the breakpoints
10 are selected from: (a) numeric values; and (b) textual values.

8. A method as in claim 3 wherein the at least one breakpoint is determined automatically based on the data.

9. A method as in claim 5 wherein the data are provided in a table,
15 wherein the coloring of an aspect of the data comprises:
coloring backgrounds of table cells according to the rules.

10. A method as in claim 1 wherein the number of groups is fewer than
20 a number of possible data values.

11. A method of operating on data, the method comprising:
providing at least one user-defined grouping rule for grouping the data into
a user-definable number of groups;

applying at least one of the grouping rules to the data to generate grouped data;

providing at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and

5 applying at least one of the scoring rules to the grouped data to score the grouped data.

12. A method of operating on data, the method comprising:

generating grouped data by applying to the data at least one user-defined
10 grouping rule for grouping the data into a user-definable number of groups; and
scoring the grouped data by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

13. A method according to claim 11 or 12 wherein the data comprises a
15 number of parameters for each of a number of cases and the scoring rule
comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the scoring of the grouped data comprises:

applying the function to the data to obtain a score for each case.
20

14. A method according to claim 13, further comprising:

sorting the scored cases by score.

15. A method according to claim 14, wherein the scored cases are
25 sorted individually.

16. A method according to claim 14, wherein the scored cases are sorted by cluster.

5 17. A system for operating on data, the system comprising:
a mechanism constructed and adapted to provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and
a mechanism constructed and adapted to apply at least one of the grouping rules to the data.

10

18. A system as in claim 17 wherein the data are provided in a table and wherein the at least one grouping rule applies to at least one user-selectable column of the table.

15

19. A system as in claim 17, wherein the at least one grouping rule defines breakpoints corresponding to the user-definable number of groups, and wherein application of the at least one rule to the data divides the data into groups based on the breakpoints.

20

20. A system as in claim 17, further comprising:
a mechanism constructed and adapted to present the grouped data in a manner that visually distinguishes the groups.

21. A system as in claim 20, wherein the grouping rules associate colors with groups and wherein the mechanism constructed and adapted to present the grouped data further comprises:

5 a mechanism constructed and adapted to color an aspect of the data according to the rules.

22. A system as in claim 20, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein
10 the mechanism constructed and adapted to present the grouped data comprises:

a mechanism constructed and adapted to color each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

15 23. A system as in any one of claims 19 and 22, wherein the breakpoints are selected from: (a) numeric values; and (b) textual values.

24. A system as in claim 19 further comprising:

20 a mechanism constructed and adapted to determine at least one breakpoint automatically, based on the data.

25. A system as in claim 21 wherein the data are provided in a table, wherein the mechanism constructed and adapted to color an aspect of the data comprises:

a mechanism constructed and adapted to color backgrounds of table cells according to the rules.

26. A system as in claim 17 wherein the number of groups is fewer
5 than a number of possible data values.

27. A system of operating on data, the system comprising:
a mechanism constructed and adapted to provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups;
10 a mechanism constructed and adapted to apply at least one of the grouping rules to the data to generate grouped data;
a mechanism constructed and adapted to provide at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and
a mechanism constructed and adapted to apply at least one of the scoring
15 rules to the grouped data to score the grouped data.

28. A system of operating on data, the system comprising:
a mechanism constructed and adapted to generate grouped data by
applying to the data at least one user-defined grouping rule for grouping the data
20 into a user-definable number of groups; and
a mechanism constructed and adapted to score the grouped data by
applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

29. A system according to claim 27 or 28 wherein the data comprises a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the mechanism constructed and adapted to score of the grouped data comprises:

a mechanism constructed and adapted to apply the function to the data to obtain a score for each case.

30. A system according to claim 29, further comprising:
a mechanism constructed and adapted to sort the scored cases by score.

31. A system according to claim 30, wherein the scored cases are sorted individually.

32. A system according to claim 30, wherein the scored cases are sorted by cluster.

33. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:

providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and
applying at least one of the grouping rules to the data.

34. A medium as in claim 33 wherein the data are provided in a table and wherein the at least one grouping rule applies to at least one user-selectable column of the table.

5 35. A medium as in claim 33 wherein the at least one grouping rule defines breakpoints corresponding to the user-definable number of groups, and wherein application of the at least one rule to the data divides the data into groups based on the breakpoints.

10 36. A medium as in claim 33, wherein the method further comprises: presenting the grouped data in a manner that visually distinguishes the groups.

15 37. A medium as in claim 36 wherein the grouping rules associate colors with groups and wherein the presenting of the grouped data further comprises: coloring an aspect of the data according to the rules.

20 38. A medium as in claim 36, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein the presenting of the grouped data comprises:

25 coloring each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

39. A medium as in any one of claims 35 and 38, wherein the breakpoints are selected from: (a) numeric values; and (b) textual values.

5 40. A medium as in claim 35 wherein the at least one breakpoint is determined automatically based on the data.

41. A medium as in claim 37 wherein the data are provided in a table, wherein the coloring of an aspect of the data comprises:
10 coloring backgrounds of table cells according to the rules.

42. A medium as in claim 33 wherein the number of groups is fewer than a number of possible data values.

15 43. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:
 providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups;
20 applying at least one of the grouping rules to the data to generate grouped data;
 providing at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and
 applying at least one of the scoring rules to the grouped data to score the
25 grouped data.

44. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:

5 generating grouped data by applying to the data at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and
 scoring the grouped data by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

10 45. A medium according to claim 43 or 44, wherein the data comprises a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the scoring of the grouped data comprises:

15 applying the function to the data to obtain a score for each case.

46. A medium according to claim 44, the method further comprising:
 sorting the scored cases by score.

20 47. A medium according to claim 46, wherein the scored cases are sorted individually.

48. A medium according to claim 46, wherein the scored cases are sorted by cluster.

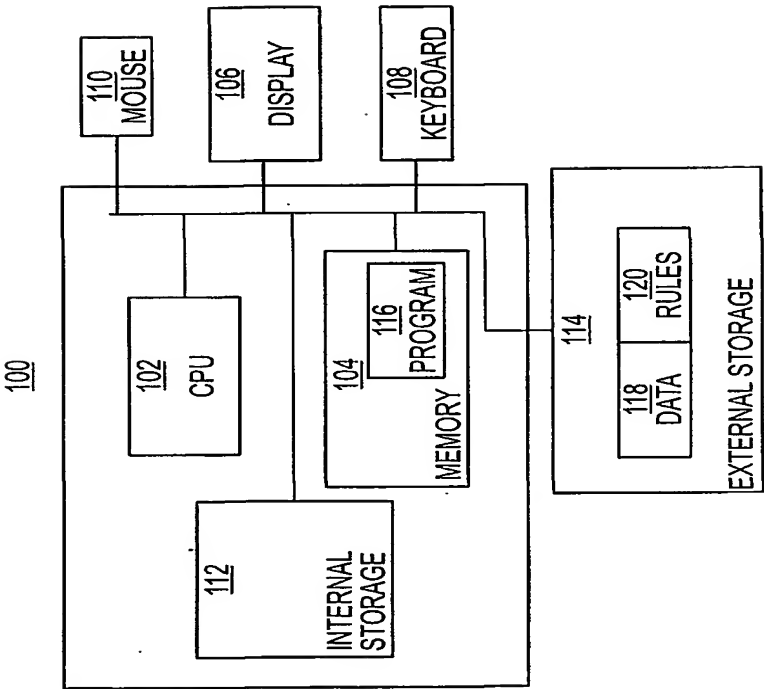


FIG. 1

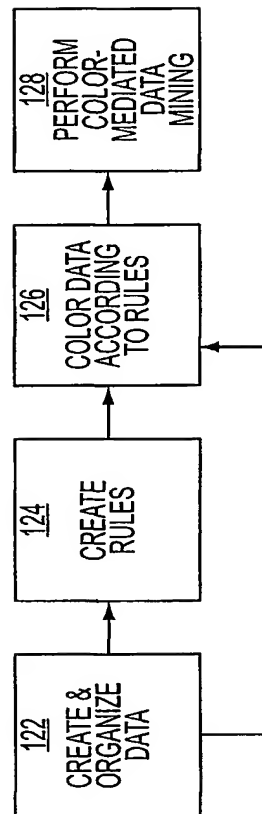


FIG. 2

Microsoft Excel - demo - PANDORA~v23.1.xls [Read Only]													
File Edit View Insert Format Tools Data Window Help													
= Cmpd													
A1	A	B	C	D	E	F	G	H	I	J	K	L	M
	Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M				
1													
2	Cmpd01	N		29	30								
3	Cmpd02	N		42	5.5	41	3	22	5				
4	Cmpd03	G		2.61	11	83	57	28	15				
5	Cmpd04	N			30	70	25	24	29				
6	Cmpd05	N		1.8	9.2	89	60	21	22				
7	Cmpd06	D	8.86	6.5	3.7	71	41	13	3				
8	Cmpd07	D	3.11	0.037	7.8	100	79	48	43				
9	Cmpd08	D		0.089	2.6	65	28	28	38				
10	Cmpd09	D	0.119			68	41	22	15				
11	Cmpd10	N	0.233			61	42	24	5				
12	Cmpd11	N	4.31			50	77	63	25				
13	Cmpd12	H	1.3	0.24		47	25	24	3				
14	Cmpd13	H	1.17	0.194	30	81	59	40	37				
15	Cmpd14	H	0.26	0.41		39	23	4	12				
16	Cmpd15	H	0.369	0.148		99	46	46	36				
17	Cmpd16	I		0.87	30	101	82	38	18				
18	Cmpd17	K		0.223	30	81	64	47	24				
19	Cmpd18	I	5.27			79	54	22	32				
20	Cmpd19	I	0.134			71	71	23	12				
21	Cmpd20	F		0.317		101	109	108	100				
22	Cmpd21	K		2.21		87	70	31	13				
23	Cmpd22	B		0.15		94	77	36	12				
24	Cmpd23	B				96	61	36	12				
25	Cmpd24	B		0.27		110	91	69	39				
26	Cmpd25	B	0.041	1.1		105	104	75	52				
27	Cmpd26	B	0.665			93	71	41	22				
28	Cmpd27	B	0.111			97	79	43	23				
29	Cmpd28	E		0.13		95	93	52	25				
30	Cmpd29	J		0.46		68	62	12	11				
31	Cmpd30	J	2.79	45		41	48	9	17				
32	Cmpd31	N											
Ready													
<div> <div> <div>DEMO 1</div> <div>DEMO 2</div> <div>DEMO 3</div> <div>clusterinfo</div> <div>DEMO</div> <div>Append Control</div> <div>Color Control</div> <div>Cluster Control</div> </div> <div> <div>302</div> <div>304</div> <div>306</div> <div>308</div> <div>310</div> <div>312</div> <div>314</div> </div> </div>													

FIG. 3A

300

File Edit View Insert Format Tools Data Window Help													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M					
1													
35	Cmpd34	J	2.27	24	30	105	62	21					
36	Cmpd35	J	0.63			44	28	18	3				
37	Cmpd36	J		0.23		93	63	42	14				
38	Cmpd37	N	12.3			58	20	38	30				
39	Cmpd38	L	0.009	0.024		90	60	51	-2				
40	Cmpd39	F	0.56	0.41		87	73	29	15				
41	Cmpd40	G	0.368	1.04		65	56	40	11				
42	Cmpd41	G	0.018	0.35		66	46	28	13				
43	Cmpd42	F	0.38	0.9		102	87	92	87				
44	Cmpd43	H	4.45	0.13		36	25	18	12				
45	Cmpd44	G	0.076	0.1		78	60	40	25				
46	Cmpd45	F	0.5			111	110	104	82				
47	Cmpd46	M	1.14			25	21	16	21				
48	Cmpd47	F	0.026			109	104	97	80				
49	Cmpd48	F	0.27			100	102	83	75				
50	Cmpd49	M	0.035	0.043		71	43	31	15				
51	Cmpd50	F	0.051			112	111	112	75				
52	Cmpd51	G	0.079			78	70	44	27				
53	Cmpd52	H	0.33	0.117	4.9	69	24	23	14				
54	Cmpd53	A	0.035	0.18	1.6	108	102	63	27				
55	Cmpd54	A	0.33			91	78	76	33				
56	Cmpd55	G	9.12			61	64	60	26				
57	Cmpd56	C	0.39			93	80	55	37				
58	Cmpd57	A	0.018			101	72	42	29				
59	Cmpd58	C	0.22			92	69	55	49				
60													
61													
62													
63													
64													
65													
66													
67													
68													
69													
70													
Ready	DEMO 1	DEMO 2	DEMO 3	clusterinfo	DEMO	Append Control	Color Control	Cluster Control					
	302	304	306	308	310	312	314						

FIG. 3B

FIG. 4A

[illegible]

SUBSTITUTE SHEET (RULE 26)

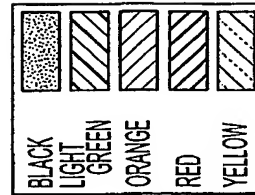


FIG. 4B

[illegible]

-312

SUBSTITUTE SHEET (RULE 26)

7/64

130

Click here to run these	
sheet	DEMO 1
column(s)	E
# of colors	6
break 1	1
break 2	5
break 3	10
break 4	
color 1	lightgreen
color 2	yellow
color 3	orange
color 4	red
Re-scale all?	no

132

134

136

138

140

142

BLACK

LIGHT GREEN

ORANGE

RED

YELLOW

FIG. 5A

SUBSTITUTE SHEET (RULE 26)

130-1

Click here to run these	
sheet	DEMO 1
column(s)	C,D
# of colors	4
break 1	0.1
break 2	1
break 3	5
break 4	
color 1	lightgreen
color 2	yellow
color 3	orange
color 4	red
Re-scale all?	no






BLACK	
LIGHT GREEN	
ORANGE	
RED	
YELLOW	

FIG. 5B

SUBSTITUTE SHEET (RULE 26)

152

144

Click here to run these	
sheet	DEMO 3
column(s)	
# of colors	6
break 1	
break 2	
break 3	
break 4	
break 5	
color 1	
color 2	
color 3	
color 4	
color 5	
color 6	
Re-scale all?	

146

148

150

FIG. 6A

10/64

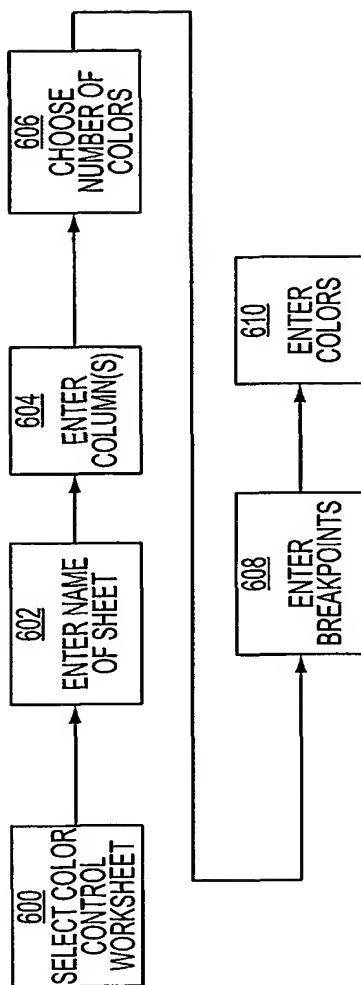


FIG. 6B

SUBSTITUTE SHEET (RULE 26)

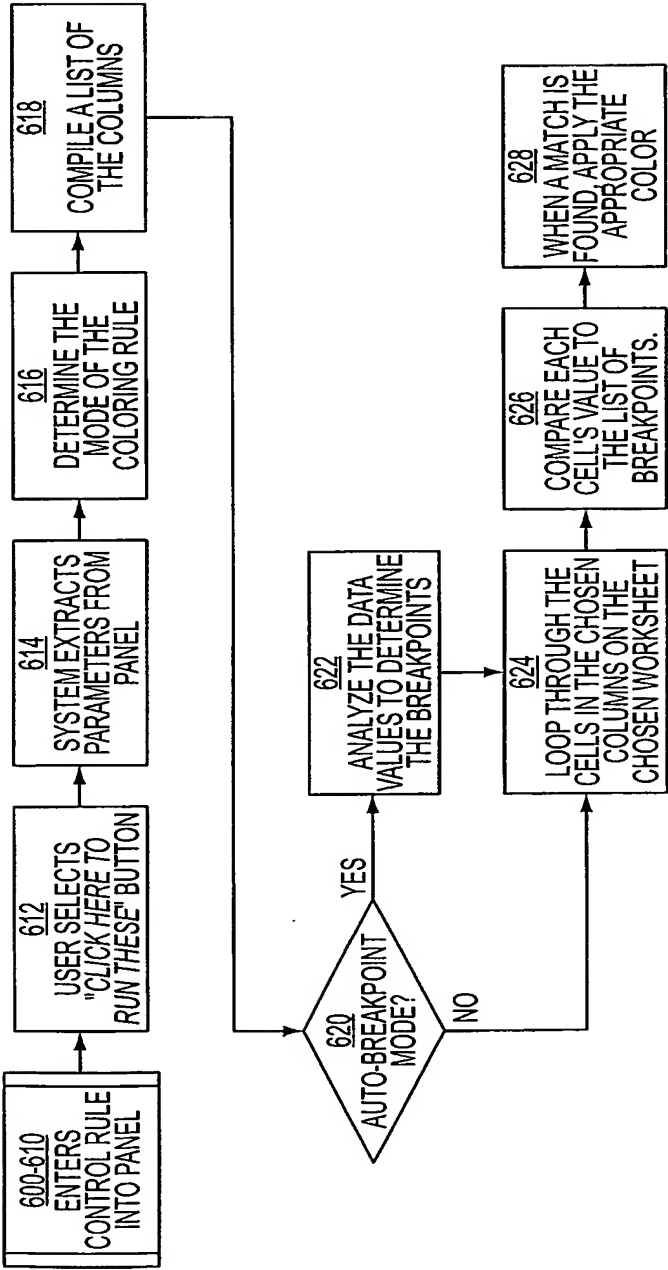


FIG. 6C

12/64

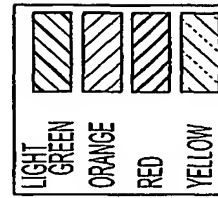


FIG. 7A

File Edit View Insert Format Tools Data Window Help																										
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O												
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M																		
1																										
2	Cmpd01	N	29	30																						
3	Cmpd02	N	42	5.5	41	3	22	5																		
4	Cmpd03	G	2.61	11	83	57	28	15																		
5	Cmpd04	N		30	70	25	24	29																		
6	Cmpd05	N	1.8	9.2	89	60	21	22																		
7	Cmpd06	D	6.5	3.7	71	41	13	3																		
8	Cmpd07	D	0.037	7.8	100	79	48	43																		
9	Cmpd08	D	0.089	2.6	65	28	28	38																		
10	Cmpd09	D	0.119		68	41	22	15																		
11	Cmpd10	N	0.233		61	42	24	5																		
12	Cmpd11	N	4.31		50	77	63	25																		
13	Cmpd12	H	1.3		47	25	24	3																		
14	Cmpd13	H	0.194	30	81	59	40	37																		
15	Cmpd14	H	0.26		39	23	4	12																		
16	Cmpd15	H	0.369		99	46	46	36																		
17	Cmpd16	I	0.87	30	101	82	38	18																		
18	Cmpd17	K	0.223	30	81	64	47	24																		
19	Cmpd18	I	5.27		79	54	22	32																		
20	Cmpd19	I	0.134		71	71	23	12																		
21	Cmpd20	F	0.317		101	109	108	100																		
22	Cmpd21	K	2.21		87	70	31	13																		
23	Cmpd22	B	0.15		94	77	36	12																		
24	Cmpd23	B			96	61	36	12																		
25	Cmpd24	B	0.27		110	91	69	39																		
26	Cmpd25	B	1.1		105	104	75	52																		
27	Cmpd26	B	0.665		93	71	41	22																		
28	Cmpd27	B	0.111		97	79	43	23																		
29	Cmpd28	E	0.13		95	93	52	25																		
30	Cmpd29	J	0.46		68	62	12	11																		
31	Cmpd30	J	45		41	48	9	17																		
32	Cmpd31	N	2.6	5.2	5	6	21	8																		
33	Cmpd32	J	2.37		62	30	15	33																		
34	Cmpd33	N			112	29	75	14																		
35	Cmpd34	J	2.27	30	105	62	62	21																		
36	Cmpd35	I	0.637		44	28	18	3																		
DEMO 1		DEMO 2	DEMO 3	clusterinfo	DEMO	Append Control	Color Control	Cluster Control																		

SUBSTITUTE SHEET (RULE 26)

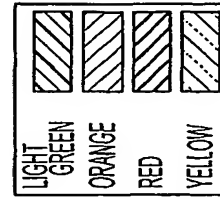


FIG. 7B

File Edit View Insert Format Tools Data Window Help													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M					
33 Cmpd32	J	2.37			62	30	15	33					
34 Cmpd33	N				112	29	75	14					
35 Cmpd34	J	2.27	24	30	105	62	62	21					
36 Cmpd35	J	0.63			44	28	18	3					
37 Cmpd36	J		0.23		93	63	42	14					
38 Cmpd37	N	12.3			58	20	38	30					
39 Cmpd38	L	0.009	0.024		90	60	51	-2					
40 Cmpd39	F	0.56	0.41		87	73	29	15					
41 Cmpd40	G	0.368	1.04		65	56	40	11					
42 Cmpd41	G	0.018	0.35		66	46	28	13					
43 Cmpd42	F	0.38	0.9		102	87	92	87					
44 Cmpd43	H	4.45	0.13		36	25	18	12					
45 Cmpd44	G	0.076	0.1		78	60	40	25					
46 Cmpd45	F	0.5			111	110	104	82					
47 Cmpd46	M	1.14			25	21	16	21					
48 Cmpd47	F	0.028			109	104	97	80					
49 Cmpd48	F	0.27			100	102	83	75					
50 Cmpd49	M	0.035	0.043		71	43	31	15					
51 Cmpd50	F	0.051			112	111	112	75					
52 Cmpd51	G	0.079			78	70	44	27					
53 Cmpd52	H	0.33	0.117	4.9	69	24	23	14					
54 Cmpd53	A	0.035	0.16	1.6	108	102	63	27					
55 Cmpd54	A	0.33			91	78	76	33					
56 Cmpd55	G	9.12			61	64	60	26					
57 Cmpd56	C	0.39			93	80	55	37					
58 Cmpd57	A	0.018			101	72	42	29					
59 Cmpd58	C	0.22			92	69	55	49					
60													
61													
62													
63													
64													
65													
66													
67													
68													
69													
70													
Cluster Control													
Color Control													
Append Control													
clusterInfo DEMO													
DEMO 3													
DEMO 2													
DEMO 1													

SUBSTITUTE SHEET (RULE 26)

14/64

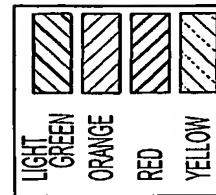


FIG. 8A

File Edit View Insert Format Tools Data Window Help														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M					
1														
2	Cmpd01	N		29	30									
3	Cmpd02	N		42	55	41	3	22	5					
4	Cmpd03	G		261	11	83	57	28	15					
5	Cmpd04	N			30	70	25	24	29					
6	Cmpd05	N		18	92	89	60	21	22					
7	Cmpd06	D	8.66	6.5	3.7	71	41	13	3					
8	Cmpd07	D	3.11	0.037	7.8	100	79	48	43					
9	Cmpd08	D		0.089	2.6	65	28	28	38					
10	Cmpd09	D	0.119			68	41	22	15					
11	Cmpd10	N	0.233			61	42	24	5					
12	Cmpd11	N	4.31			50	77	63	25					
13	Cmpd12	H	1.3	0.24		47	25	24	3					
14	Cmpd13	H	1.17	0.194	30	81	59	40	37					
15	Cmpd14	H	0.26	0.41		39	23	4	12					
16	Cmpd15	H	0.369	0.148		99	46	46	36					
17	Cmpd16	I		0.87	30	101	82	38	18					
18	Cmpd17	K		0.223	30	81	84	47	24					
19	Cmpd18	I	5.27			79	54	22	32					
20	Cmpd19	I	0.134			71	71	23	12					
21	Cmpd20	F		0.317		101	109	108	100					
22	Cmpd21	K		2.21		87	70	31	13					
23	Cmpd22	B		0.15		94	77	36	12					
24	Cmpd23	B				96	61	36	12					
25	Cmpd24	B		0.27		110	91	69	39					
26	Cmpd25	B	0.041	1.1		105	104	75	52					
27	Cmpd26	B	0.665			93	71	41	22					
28	Cmpd27	B	0.111			97	79	43	23					
29	Cmpd28	E		0.13		95	93	52	25					
30	Cmpd29	J		0.46		68	82	12	11					
31	Cmpd30	J	2.79	45		41	48	9	17					
32	Cmpd31	N		2.6	52	5	6	21	8					
33	Cmpd32	J	2.37			62	30	15	33					
34	Cmpd33	N				112	29	75	14					
Ready														
DEMO 1 / DEMO 2 / DEMO 3 / clusterinfo DEMO / Append Control / Color Control / Cluster Control / NUM														

SUBSTITUTE SHEET (RULE 26)

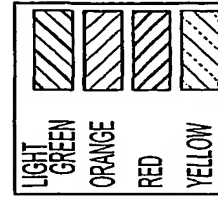


FIG. 8B

File Edit View Insert Format Tools Data Window Help														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M						
1														
35	Cmpd34	J	2.27	24	30	62	82	21						
36	Cmpd35	J	0.63			28	18	3						
37	Cmpd36	J		0.23		63	42	14						
38	Cmpd37	N	12.3			20	38	30						
39	Cmpd38	L	0.009	0.024		60	51	2						
40	Cmpd39	F	0.58	0.41		73	29	15						
41	Cmpd40	G	0.368	1.04		56	40	11						
42	Cmpd41	G	0.018	0.35		46	28	13						
43	Cmpd42	F	0.38	0.9		87	92	87						
44	Cmpd43	H	4.45	0.13		25	18	12						
45	Cmpd44	G	0.076	0.1		60	40	25						
46	Cmpd45	F	0.5			110	104	82						
47	Cmpd46	M	1.14			21	16	21						
48	Cmpd47	F	0.026			109	104	80						
49	Cmpd48	F	0.27			100	83	75						
50	Cmpd49	M	0.035	0.043		71	43	15						
51	Cmpd50	F	0.051			112	111	75						
52	Cmpd51	G	0.079			78	70	27						
53	Cmpd52	H	0.33	0.117	4.9	69	23	14						
54	Cmpd53	A	0.035	0.18	1.6	108	102	27						
55	Cmpd54	A	0.33			81	78	33						
56	Cmpd55	G	9.12			61	64	26						
57	Cmpd56	C	0.39			93	80	37						
58	Cmpd57	A	0.018			101	72	29						
59	Cmpd58	C	0.22			92	69	49						
60														
61														
62														
63														
64														
65														
66														
67														
68														
69														
70														
71														
Ready														
<div> <div> <div>DEMO 1</div> <div>DEMO 2</div> <div>DEMO 3</div> <div>clusterinfo DEMO</div> <div>Append Control</div> <div>Color Control</div> <div>Cluster Control</div> </div> <div>14</div> <div>NUM</div> </div>														

SUBSTITUTE SHEET (RULE 26)

[illegible]

SUBSTITUTE SHEET (RULE 26)

[illegible]

FIG. 9B

SUBSTITUTE SHEET (RULE 26)

19/64

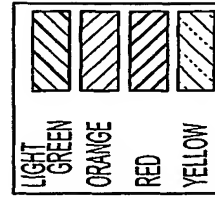


FIG. 10B

Microsoft Excel - demo "PANDORA~v23.1.xls [Read-Only]"																	
File Edit View Insert Format Tools Data Window Help																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
34	F2_A	F2	A	32													
35	F2_B	F2	B	61													
36	F2_C	F2	C	72													
37	F2_D	F2	D	53													
38	F3_A	F3	A	14													
39	F3_B	F3	B	38													
40	F3_C	F3	C	33													
41	F3_D	F3	D	42													
42	F4_A	F4	A	41													
43	F4_B	F4	B	9													
44	F4_C	F4	C	18													
45	F4_D	F4	D	22													
46	F5_A	F5	A	1													
47	F5_B	F5	B	21													
48	F5_C	F5	C	23													
49	F5_D	F5	D	22													
50	F6_A	F6	A	49													
51	F6_B	F6	B	35													
52	F6_C	F6	C	26													
53	F6_D	F6	D	14													
54	F7_A	F7	A	4													
55	F7_B	F7	B	19													
56	F7_C	F7	C	9													
57	F7_D	F7	D	15													
58																	
59																	
60																	
61																	
62																	
63																	
64																	
65																	
66																	
67																	
68																	
69																	
70																	
71																	
72																	
73																	
74																	
75																	

SUBSTITUTE SHEET (RULE 26)

20/64

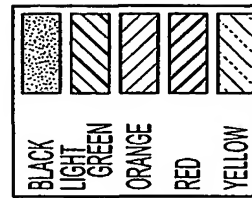


FIG. 11A

File Edit View Insert Format Tools Data Window Help

TIP: To quickly navigate among the worksheets in your workbook, especially if you have a lot of them, RIGHT-click on any of the tab-scrolling arrow buttons at the lower left of the screen, to get a list of sheet names to pick from.

Other shortcuts (see also the PANDORA menu)
Ctrl+Shift+J goes to the Append Control

HELP
Clear all entries on this sheet

© Zeneca Inc. 1998

TIP: To capture the name of a sheet for an entry into one of the control panels, double-click the sheet's tab to get a "Rename Sheet" dialog box. Then hit CTRL-C to Edit: Copy the name to the clipboard, and click Cancel on the Rename box. Go to the cell where you want to paste the sheet name, and either hit CTRL-V or do an Edit: Paste.

Name:	Series	Name:	Score
Enlarge Cluster Starts	Demo 1	Enlarge Cluster Starts	
Sheet #	Demo 1	Sheet #	Demo 3
Cluster Col	B	Cluster Col	
Shrink Cluster Starts		Shrink Cluster Starts	

Color	Score
red	-1
orange	0
yellow	1
lightgreen	2

Score and Sort Clusters	Column(s)	Rel. Weight
C,D	1	1
E	1	1

Score Dose-Response	Columns	F:

Name:	Series	Name:	Score
Enlarge Cluster Starts	amines	Enlarge Cluster Starts	
Sheet #	Demo 2	Sheet #	Demo 3
Cluster Col	C	Cluster Col	
Shrink Cluster Starts		Shrink Cluster Starts	

Color	Score
red	0
orange	1
yellow	2

Score and Sort Clusters	Column(s)	Rel. Weight
C,D	1	1
E	1	1

Score Dose-Response	Columns	F:

Name:	Series	Name:	Score
Enlarge Cluster Starts	adds	Enlarge Cluster Starts	
Sheet #	Demo 2	Sheet #	Demo 3
Cluster Col	B	Cluster Col	
Shrink Cluster Starts		Shrink Cluster Starts	

Color	Score
red	0
orange	1
yellow	2

Score and Sort Clusters	Column(s)	Rel. Weight
C,D	1	1
E	1	1

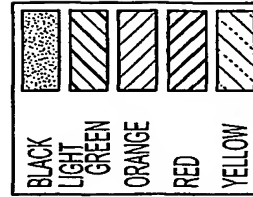
Score Dose-Response	Columns	F:

Append Control / Cluster Control / 14

SUBSTITUTE SHEET (RULE 26)

314

1100



TIP: To capture the name of a sheet for an entry into one of the control panels, double-click the sheet's tab to get a "Rename Sheet" dialog box. Then hit CTRL-C to Edit: Copy the name to the clipboard, and click Cancel on the Rename box. Go to the cell where you want to paste the sheet name, and either hit CTRL-V or do an Edit: Paste.

File	Edit	View	Insert	Format	Tools	Data	Window	Help
<div>Score Dose-Response Columns F:1</div> <div>Name: acids Enlarge Cluster Starts Sheet # Demo 2 Cluster Col B Shrink Cluster Starts</div> <div>Color Score red 0 orange 1 yellow 2 lightgreen 3</div> <div>Score and Sort Clusters Columns(s) Rel. Weight D 1</div> <div>Score Dose-Response Columns a:a</div>								
<div>Score Dose-Response Columns F:1</div> <div>Name: amines Enlarge Cluster Starts Sheet # Demo 2 Cluster Col C Shrink Cluster Starts</div> <div>Color Score red 0 orange 1 yellow 2 lightgreen 3</div> <div>Score and Sort Clusters Columns(s) Rel. Weight D 1</div> <div>Score Dose-Response Columns F:1</div>								
<div>Score Dose-Response Columns</div> <div>Name: amine Enlarge Cluster Starts Sheet # Demo 3 Cluster Col Shrink Cluster Starts</div> <div>Color Score</div> <div>Score and Sort Clusters Columns(s) Rel. Weight</div> <div>Score Dose-Response Columns F:1</div>								
<div>Score Dose-Response Columns</div> <div>Name: amine Enlarge Cluster Starts Sheet # Demo 3 Cluster Col Shrink Cluster Starts</div> <div>Color Score</div> <div>Score and Sort Clusters Columns(s) Rel. Weight</div> <div>Score Dose-Response Columns F:1</div>								

1 4 14 DEMO 1 DEMO 2 DEMO 3 clusterinfo DEMO Append Control Color Control Cluster Control 4

SUBSTITUTE SHEET (RULE 26)

FIG. 11B

314

Figure 1 is a schematic diagram of a data processing system 1100. The system includes a Name field 1102, an Enlarge Cluster Starts field 1104, a Sheet # field 1106, a Cluster Col field 1110, and a Shrink Cluster Starts field 1112. The Cluster Col field 1110 is divided into four sub-fields: red 1112, orange 1114, yellow 1116, and lightgreen 1118. The Shrink Cluster Starts field 1112 is divided into four sub-fields: 0 1112, 1 1114, 2 1116, and 3 1118. The system also includes a Score and Sort Clusters field 1108, a Column(s) field 1110, and a Rel. Weight field 1114. A legend 1116 shows the color coding for the Cluster Col field: Black, Light Green, Orange, Red, and Yellow.

FIG. 11C

SUBSTITUTE SHEET (RULE 26)

[illegible]

FIG. 11D

SUBSTITUTE SHEET (RULE 26)

24/64

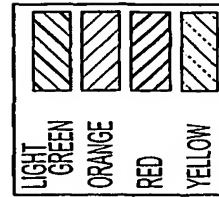


FIG. 12

Microsoft Excel - demo "PANDORA"~v23.1.xls [Read-Only]									
File Edit View Insert Format Tools Data Window Help									
B1 = acid									
	A	B	C	D	E	F	G	H	I
1	mixture	acid	amine group	activity @ 1uM					
2	E1_A	E1	A	47					
3	E1_B	E1	B	38					
4	E1_C	E1	C	40					
5	E1_D	E1	D	32					
6	E2_A	E2	A	7					
7	E2_B	E2	B	69					
8	E2_C	E2	C	40					
9	E2_D	E2	D	17					
10	E3_A	E3	A	17					
11	E3_B	E3	B	21					
12	E3_C	E3	C	20					
13	E3_D	E3	D	14					
14	E4_A	E4	A	22					
15	E4_B	E4	B	17					
16	E4_C	E4	C	27					
17	E4_D	E4	D	12					
18	E5_A	E5	A	3					
19	E5_B	E5	B	5					
20	E5_C	E5	C	28					
21	E5_D	E5	D	10					
22	E6_A	E6	A	5					
clusterinfo DEMO clusterinfo DEMO 3 DEMO 2 DEMO 1 DEMO 0									
Ready									

SUBSTITUTE SHEET (RULE 26)

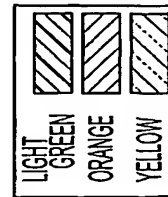


FIG. 13A

File Edit View Insert Format Tools Data Window Help														
CMPD ID	# of replicates	S#	#PTS Test1	IC50 (nM) Test1	DATE Test1	#PTS Test2	IC50 (nM) Test2	DATE Test2	SELECTIVITY	K	L	M	N	
1														
2	(-)pentazocine	4	4.25	no effect >10000	7/20/97	4	8384	4/10/97						
3	(+)pentazocine	4	15	no effect >10000	7/20/97	4	>10000 (blank)	4/10/97						
4	(+/-)pentazocine	4	4.25	8859	7/20/97	4	3133	35530	0.00					
5	0277	3	3	165.3333333	3/6/97	3	(blank) >10000	35498	0.00					
6	0278	8	1.5	52.75	4/3/97	3	3575	35498	38.21					
7	0629	1	4	no effect	10/9/97									
8	0681	3	1	158.8666667	3/6/97	3	4100	3/9/97	15.59					
9	0697	1	4	no effect	10/6/97									
10	0793	10	1.4	32.5	4/17/97	4.285714286	26885	35539.85714	679.87					
11	0858	1	2	40	8/13/97	5	1283	8/18/97	32.08					
12	1058	1	3	445	6/19/97	4	4000	6/23/97	8.99					
13	10817-018-C	1	5	no effect	1/22/98									
14	1210	3	1	36	6/19/97	6	75000	6/24/97	2093.23					
15	1329	1	2	20	8/13/97	5	23161	8/18/97	1158.05					
16	1336	1	1	106	8/6/97	5	32903	8/7/97	310.41					
17	1337	1	1	>300	8/14/97	5	43105	8/20/97						
18	1338	1	1	no effect	8/6/97	5	>100000	8/7/97						
19	1339	1	1	293	8/14/97	5	85442	8/19/97	291.51					
20	1341	1	1	no effect	8/14/97	5	>100000	8/19/97						
21	1342	1	1	141	8/6/97	5	34184	8/7/97	242.44					
22	1343	1	1	149	7/29/97	5	62285	7/31/97	418.09					
23	1344	1	1	145	7/29/97	5	5080	7/30/97	35.03					
24	1362	1	1	225	7/19/97	4	53000	7/29/97	235.56					
25	1369	1	1	>300	7/19/97	4	no effect	7/29/97						
26	1420	1	1	111	8/13/97	5	4657	8/18/97	41.95					
27	1431	1	1	no effect	7/29/97	5	16967	7/30/97						
28	1439	1	1	no effect	8/14/97	5	no effect	8/20/97						
29	1444	1	1	269	8/14/97	5	>100000	8/20/97						
30	1445	1	1	no effect	10/8/97									
31	1446	1	1	no effect	8/14/97	5	no effect	8/20/97						
32	1447	1	1	no effect	10/8/97									
33	1453	1	1	33	7/25/97	5	15484	7/30/97	469.21					
34	1463	1	1	>300	9/25/97									
35	1464	1	1	no effect	10/3/97									
36	1465	1	1	no effect	10/8/97									
37	1466	1	1	no effect	8/15/97	5	>100000	8/21/97						
38	1467	1	1	no effect	9/18/97									
39	1468	1	1	no effect	02/25/97									
40	1469	1	1	no effect	02/25/97									
<div> <div>Ready</div> <div> <div>DEMO 1</div> <div>DEMO 2</div> <div>DEMO 3</div> </div> <div> <div>Append Control</div> <div>Color Control</div> <div>Cluster Control</div> </div> <div>NUM</div> </div>														

SUBSTITUTE SHEET (RULE 26)

FIG. 13B

File Edit View Insert Format Tools Data Window Help														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	
CMPD ID	# of replicates	S#	#PTS Test1	IC50 (nM) Test1	DATE Test1	#PTS Test2	IC50 (nM) Test2	DATE Test2	SELECTIVITY					
1	1	2	4	134	6/20/97	4	11000	6/25/97	82.09					
294	1	1	5	35	6/20/97	5	6000	6/25/97	171.43					
295	1	3	5	54	9/17/97	5	2660	9/16/97	49.26					
296	1	3	5	>300	9/4/97	5	2486	9/1/97						
297	1	6	5	56	9/17/97	5	1330	9/16/97	23.75					
298	1	1	5	10	6/18/97	5	10000	6/23/97	1000.00					
299	1	2	4	>300	6/18/97	4	>100000	6/23/97						
300	1	2	3	no effect	10/6/97									
301	1	8	5	>10000	6/12/97	4	419	4/10/97						
302	1	1	5	82	6/26/97	5	19000	7/3/97	231.71					
303	1	8	5	63	6/12/97	5	200	1/27/97	3.71					
304	1	1	5	185	7/12/97	5	2706	7/16/97	14.63					
305	1	2	3	no effect	9/25/97									
306	1	3	5	165	8/15/97	5	>100000	8/21/97						
307	1	1	5	59	8/13/97	5	1872	8/18/97	31.73					
308	1	1	5	39	8/13/97	5	2836	8/18/97	72.72					
309	1	3	5	>300	7/29/97	5	39852	7/31/97						
310	1	2	5	>300	9/18/97	5	no effect	9/16/97						
311	1	2	5	213	9/4/97	5	3233	9/9/97	15.18					
312	1	2	5	no effect	9/4/97	5	9480	9/9/97						
313	1	2	5	no effect	9/4/97	5	10830	9/9/97						
314	1	3	5	106	7/28/97	5	16000	7/10/97	150.94					
315	1	1	5	44	8/13/97	5	2013	8/18/97	45.75					
316	1	1	5	14	8/13/97	5	1168	8/18/97	83.43					
317	1	2	5	no effect	9/3/97	5	3143	9/9/97	0.00					
318	1	1	5	23	1/22/97									
319	1	3	3	>300	10/3/97									
320	3	2	5	23,333,333	11/14/97	5	1880	11/19/97	82.53					
321	1	2	3	no effect	10/8/97									
322	2		5	no effect	8/24/97	5	no effect, (blank)	8/20/97						
323	2	4	4	>10000	3/5/97	5	21360	2/2/97						
324	1	5	5	722	6/12/97	5	800	2/5/97	1.11					
325	1	5	5	29	9/17/97	5	3570	9/16/97	123.10					
326	1	4	4	>10000	3/7/97	5	75170	2/6/97						
327	2	5	5	no effect > 10000	8/24/97	5	no effect, (blank)	8/20/97						
328	2	5	5	53555	8/24/97	5	(blank), no effect	8/20/97	0.00					
329	2													
330	2													
331	2													
DEMO 1 / DEMO 2 / DEMO 3 / clusterinfo DEMO / Append Control / Color Control / Cluster Control / 14														
Ready														

SUBSTITUTE SHEET (RULE 26)

27/64

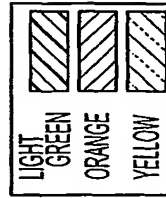


FIG. 13C

A	B	C	D	E	F	G	H	I	J	K	L	M	N
CMPD ID	# of replicates	S#	#PTS Test1	IC50 (nM) Test1	DATE Test1	#PTS Test2	IC50 (nM) Test2	DATE Test2	SELECTIVITY				
1													
2	(-)pentazocine	4	4.25	no effect > 10000	7/20/97	4	8384	4/10/97					
3	(+)pentazocine	4	5	no effect > 10000	7/20/97	4	> 10000 (blank)	4/10/97					
4	(+/-)pentazocine	4	4.25	8859	7/20/97	4	3133	35530	0.00				
5	0277	3	3	165.3333333	3/6/97	3	(blank) > 10000	35498	0.00				
6	0278	8	1.5	52.75	4/3/97	3	3575	35498	38.21				
7	0629	1	4	no effect	10/9/97								
8	0681	3	1	158.6666667	3/6/97	3	4100	3/9/97	15.59				
9	0697	1	4	no effect	10/6/97								
10	0793	10	1.4	32.5	4/17/97	4	2885	35539.85714	678.87				
11	0958	1	2	40	8/13/97	5	1283	8/18/97	32.08				
12	1058	1	3	445	6/18/97	4	4000	6/23/97	8.99				
13	10817-018-C	1	5	no effect	1/22/98								
14	1210	3	1										
15	1329	1	2										
16	1336	1	1	5									
17	1337	1	1	5									
18	1338	1	1	5									
19	1339	1	1	5									
20	1341	1	1	5									
21	1342	1	1	5									
22	1343	1	1	5									
23	1344	1	1	5									
24	1382	1	1	4									
25	1389	1	1	4									
26	1420	1	1	5									
27	1431	1	1	5									
28	1439	1	1	5									
29	1444	1	1	5									
30	1445	1	1	3									
31	1446	1	1	5									
32	1447	1	1	3									
33	1453	1	1	5									
34	1463	1	1	3									
35	1464	1	1	3									
36	1465	1	1	3									
37	1466	1	1	5									
38	1467	1	1	5									
39	1468	1	1	5									
40	1469	1	1	5									

ENTER SCALING FACTOR FOR DISPLAY

Enter a vertical scale factor, relative to current size:

>1 for enlargement
<1 for reduction
-1 for auto-fit (restore to "normal")

0.1

OK Cancel

318

306

Ready

Cluster Control

Color Control

Append Control

DEMO 3

DEMO 2

DEMO 1

NUM

SUBSTITUTE SHEET (RULE 26)

28/64

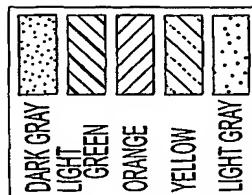


FIG. 13D

[illegible]

SUBSTITUTE SHEET (RULE 26)

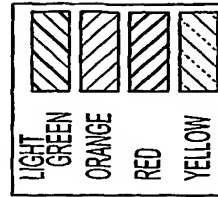


FIG. 14A

File Edit View Insert Format Tools Data Window Help									
A	B	C	D	E	F	G	H	I	J
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M	
1									
2	Cmpd01	N		30				5	
3	Cmpd02	N		42	41	3	22	5	
4	Cmpd03	G		5.5					
5	Cmpd04	N		2.61	63	57	28	15	
6	Cmpd05	N							
7	Cmpd06	D							
8	Cmpd07	D							
9	Cmpd08	D							
10	Cmpd09	D							
11	Cmpd10	N							
12	Cmpd11	N							
13	Cmpd12	H							
14	Cmpd13	H							
15	Cmpd14	H							
16	Cmpd15	H							
17	Cmpd16	I							
18	Cmpd17	K							
19	Cmpd18	I							
20	Cmpd19	I							
21	Cmpd20	F							
22	Cmpd21	K							
23	Cmpd22	B							
24	Cmpd23	B			95	61	36	12	
25	Cmpd24	B		0.27	110	91	69	39	
26	Cmpd25	B	0.041	1.1	105	104	75	52	
27	Cmpd26	B	0.665		93	71	41	22	
28	Cmpd27	B	0.111		97	79	43	23	
29	Cmpd28	E		0.13	95	93	52	25	
30	Cmpd29	J		0.46	68	62	12	11	
31	Cmpd30	J	2.79	45	41	48	9	17	
32	Cmpd31	N		5.2	5	6	21	8	
33	Cmpd32	J	2.37	2.6	62	30	15	33	
34	Cmpd33	N			112	29	75	14	
					<div>OK</div> <div>Cancel</div>				
<p>SCORING OF UNCOLORED CELLS</p> <p>PANDORA has found uncolored cells (color="none") in the specified columns.</p> <p>How do you want uncolored cells to be scored for sorting?</p> <p><input checked="" type="radio"/> Use score for the color "none" on my list.</p> <p>***THIS IS RECOMMENDED***</p> <p>You should enter the color "none" on your list on the Cluster Control sheet with an appropriate score relative to the others.</p> <p><input type="radio"/> uncolored = AVERAGE of colors in same row</p> <p>***WARNING***</p> <p>This mode should only be used with great caution. See the manual.</p>									
<p>1402</p>									
<p>314</p>									
<p>302</p>									
<p>For Help on dialog settings, press F1</p>									
<p>DEMO 1 DEMO 2 DEMO 3 clusterInfo DEMO Append Control Color Control Cluster Control</p>									
<p>SUM = 0</p>									

SUBSTITUTE SHEET (RULE 26)

FIG. 14B

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M						
1															
2	Cmpd38	L	0.009	0.024		90	60	51	-2						
3	Cmpd44	G	0.076	0.1		78	60	40	25						
4	Cmpd49	M	0.035	0.043		71	43	31	15						
5	Cmpd53	A	0.035	0.18	1.6	108	102	63	27						
6	Cmpd08	D		0.089	2.6	65	28	28	38						
7	Cmpd41	G	0.018	0.35		66	46	28	13						
8	Cmpd52	H	0.33	0.117	4.9	69	24	23	14						
9	Cmpd07	D	3.11	0.037	7.8	100	79	48	43						
10	Cmpd14	H	0.26	0.41		39	23	4	12						
11	Cmpd15	H	0.389	0.148											
12	Cmpd25	B	0.041	1.1											
13	Cmpd39	F	0.56	0.41											
14	Cmpd42	F	0.38	0.9											
15	Cmpd47	F	0.026												
16	Cmpd50	F	0.051												
17	Cmpd51	G	0.079												
18	Cmpd57	A	0.018												
19	Cmpd09	D	0.119												
20	Cmpd10	N	0.233												
21	Cmpd12	H	1.3	0.24											
22	Cmpd19	I	0.134												
23	Cmpd20	F		0.317											
24	Cmpd22	B		0.15											
25	Cmpd24	B		0.27		110	91	69	39						
26	Cmpd26	B	0.665			93	71	41	22						
27	Cmpd27	B	0.111			97	79	43	23						
28	Cmpd28	E		0.13		95	93	62	25						
29	Cmpd29	J		0.46		68	62	12	11						
30	Cmpd35	J	0.63			44	28	18	3						
31	Cmpd36	J		0.23		93	63	42	14						
32	Cmpd40	G	0.368	1.04		65	56	40	11						
33	Cmpd43	H	4.45	0.13		36	25	18	12						
34	Cmpd45	F	0.5			111	110	104	82						

Microsoft Excel

DONE!

Summary: 58 rows in 58 clusters
3 columns used in scoring
Elapsed time: 5 seconds

Two sheets have been added or overwritten in your workbook:
"DEMO 1 SCORES by Cmpd"
"DEMO 1 SORTED by Cmpd un=ze"

[OK]

SUBSTITUTE SHEET (RULE 26)

File Edit View Insert Format Tools Data Window Help															
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
cluster label	# of cmpds	score max. 100 (uncolored = zero)	score max. 100 (uncolored = average)	← Scoring of sheet "DEMO 1" using parameter set "Cmpd"											
1															
2	Cmpd38	1	67	100											
3	Cmpd44	1	67	100											
4	Cmpd49	1	67	100											
5	Cmpd53	1	67	67											
6	Cmpd08	1	50	75											
7	Cmpd41	1	50	75											
8	Cmpd52	1	50	50											
9	Cmpd07	1	33	33											
10	Cmpd14	1	33	50											
11	Cmpd15	1	33	50											
12	Cmpd25	1	33	50											
13	Cmpd39	1	33	50											
14	Cmpd42	1	33	50											
15	Cmpd47	1	33	100											
16	Cmpd50	1	33	100											
17	Cmpd51	1	33	100											
18	Cmpd57	1	33	100											
19	Cmpd09	1	17	50											
20	Cmpd10	1	17	50											
21	Cmpd12	1	17	25											
22	Cmpd19	1	17	50											
23	Cmpd20	1	17	50											
24	Cmpd22	1	17	50											
25	Cmpd24	1	17	50											
26	Cmpd26	1	17	50											
27	Cmpd27	1	17	50											
28	Cmpd28	1	17	50											
29	Cmpd29	1	17	50											
30	Cmpd35	1	17	50											
31	Cmpd36	1	17	50											
32	Cmpd40	1	17	25											
33	Cmpd43	1	17	25											
34	Cmpd45	1	17	50											
35	Cmpd48	1	17	50											
36	Cmpd54	1	17	50											
37	Cmpd56	1	17	50											
Ready															
DEMO 1 SCORES by Cmpd															
DEMO 1 SORTED by Cmpd un=ze / DEMO 2 / DEMO 3 / clusterid															
NUM															

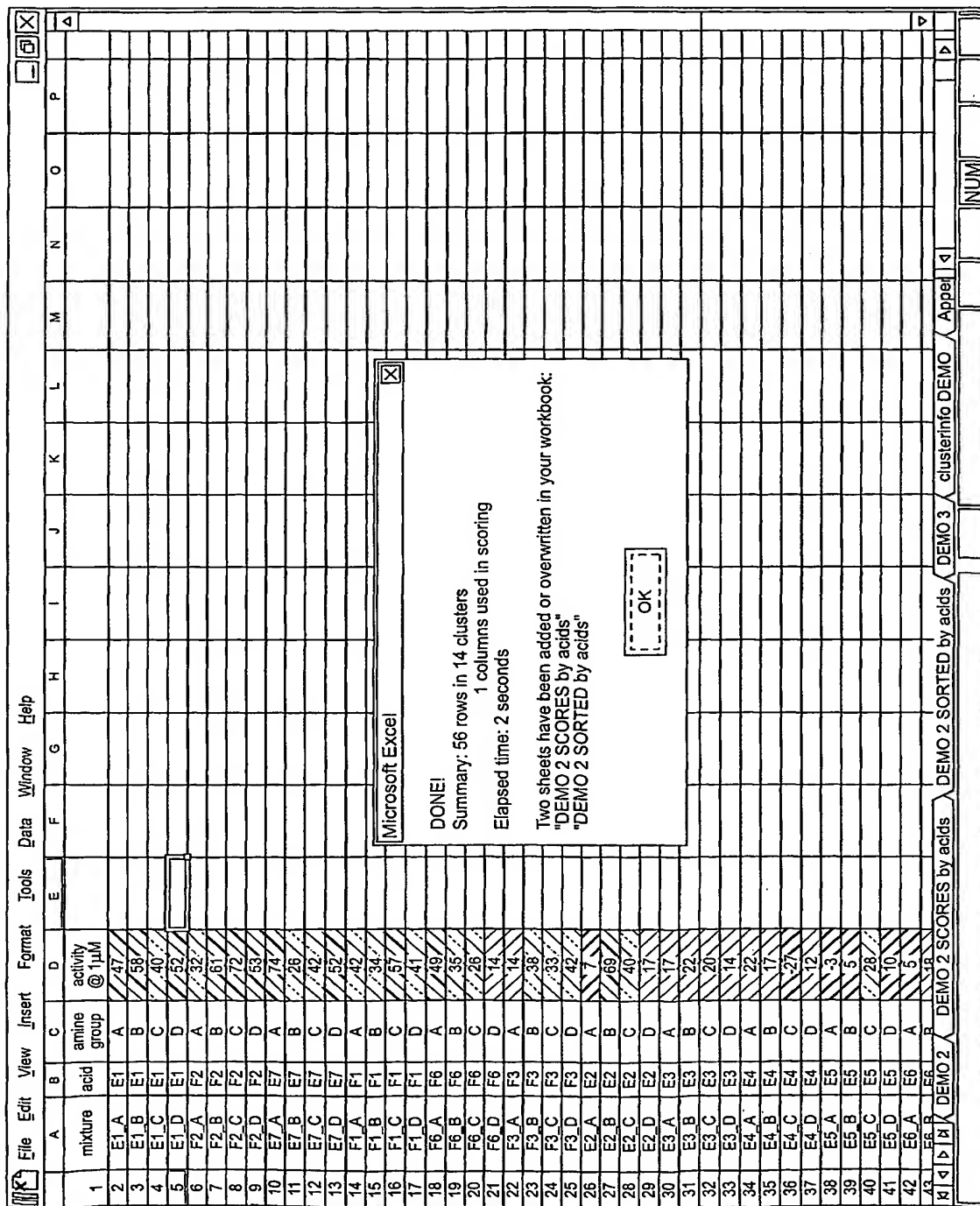
SUBSTITUTE SHEET (RULE 26)

File Edit View Insert Format Tools Data Window Help														A	
30	Cmpd35	1	17	50											
31	Cmpd36	1	17	50											
32	Cmpd40	1	17	25											
33	Cmpd43	1	17	25											
34	Cmpd45	1	17	50											
35	Cmpd48	1	17	50											
36	Cmpd54	1	17	50											
37	Cmpd56	1	17	50											
38	Cmpd58	1	17	50											
39	Cmpd05	1	0	0											
40	Cmpd11	1	0	0											
41	Cmpd13	1	0	0											
42	Cmpd16	1	0	0											
43	Cmpd17	1	0	0											
44	Cmpd21	1	0	0											
45	Cmpd23	1	0	0											
46	Cmpd31	1	0	0											
47	Cmpd32	1	0	0											
48	Cmpd33	1	0	0											
49	Cmpd46	1	0	0											
50	Cmpd02	1	-33	-50											
51	Cmpd03	1	-33	-50											
52	Cmpd04	1	-33	-100											
53	Cmpd06	1	-33	-33											
54	Cmpd18	1	-33	-100											
55	Cmpd30	1	-33	-50											
56	Cmpd37	1	-33	-100											
57	Cmpd55	1	-33	-100											
58	Cmpd01	1	-67	-100											
59	Cmpd34	1	-67	-67											
60															
61															
62															
63															
64															
65															
66															
67															
68															
69															
70															
71															
DEMO 1 DEMO 2 DEMO 3 cluster															

SUBSTITUTE SHEET (RULE 26)

File Edit View Insert Format Tools Data Window Help															4	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	cluster label	# of cmpds	score (uncolored = zero)	score max. 100 (uncolored = average)	← Scoring of sheet "DEMO 2" using parameter set "acids"											
2	E1	4	92	92												
3	F2	4	92	92												
4	E7	4	83	83												
5	F1	4	75	75												
6	F6	4	67	67												
7	F3	4	58	58												
8	E2	4	50	50												
9	E3	4	33	33												
10	E4	4	25	25												
11	E5	4	17	17												
12	E6	4	17	17												
13	F4	4	17	17												
14	F5	4	8	8												
15	F7	4	8	8												
16																
17																
18																
19																
20																
21																
22																
23																
24																
25																
26																
27																
28																
29																
30																
31																
32																
33																
34																
35																
36																
37																
38																
39																
K1		DEMO 2	DEMO 2 SCORES by acids		DEMO 2 SORTED by acids		DEMO 3		clusterinfo DEMO		Appet		4			

SUBSTITUTE SHEET (RULE 26)



SUBSTITUTE SHEET (RULE 26)

FIG. 15B

	A	B	C	D	E
	Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M
1					
2	Cmpd01	90	22	16	19
3	Cmpd02	41	3	22	5
4	Cmpd03	83	57	28	15
5	Cmpd04	70	25	24	29
6	Cmpd05	89	60	21	22
7	Cmpd06	71	41	13	3
8	Cmpd07	100	79	48	43
9	Cmpd08	65	28	28	38
10	Cmpd09	68	41	22	15
11	Cmpd10	61	42	24	5
12	Cmpd11	50	77	63	25
13	Cmpd12	47	25	24	3
14	Cmpd13	81	59	40	37
15	Cmpd14	39	23	4	12
16	Cmpd15	99	46	46	36
17	Cmpd16	101	82	38	18
18	Cmpd17	81	64	47	24
19	Cmpd18	79	54	22	32
20	Cmpd19	71	71	23	12
21	Cmpd20	101	109	108	100

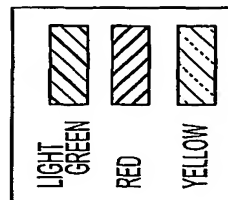


FIG. 16A

SUBSTITUTE SHEET (RULE 26)

36/64

Click here to run these	
sheet	DEMO 1
column(s)	B:E
# of colors	3
break 1	33
break 2	67
break 3	
color 1	red
color 2	yellow
color 3	lightgreen
Re-scale all?	





BLACK	
LIGHT GREEN	
RED	
YELLOW	

FIG. 16B

37/64

	A	B	C	D	E	F	G	H
	Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10-6M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10-6M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10-7M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10-7M 1.00e-07	D-R -iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd
1								
2	Cmpd20	101	109	108	100	75	100	100
3	Cmpd07	100	79	48	43	83	77	79
4	Cmpd13	81	59	40	37	83	70	75
5	Cmpd15	99	46	46	36	83	70	75
6	Cmpd16	101	82	38	18	92	50	70
7	Cmpd17	81	64	47	24	92	43	67
8	Cmpd03	83	57	28	15	92	23	57
9	Cmpd05	89	60	21	22	92	23	57
10	Cmpd06	71	41	13	3	92	23	57
11	Cmpd09	68	41	22	15	92	23	57
12	Cmpd11	50	77	63	25	67	47	57
13	Cmpd18	79	54	22	32	92	23	57
14	Cmpd19	71	71	23	12	83	30	56
15	Cmpd10	61	42	24	5	83	20	51
16	Cmpd01	90	22	16	19	83	10	46
17	Cmpd04	70	25	24	29	83	10	46
18	Cmpd02	41	3	22	5	83	7	45
19	Cmpd08	65	28	28	38	58	33	45
20	Cmpd12	47	25	24	3	83	7	45
21	Cmpd14	39	23	4	12	83	7	45

FIG. 16C



SUBSTITUTE SHEET (RULE 26)

38/64

	A	B	C	D	E	F	G	H
	Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10-6M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10-6M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10-7M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10-7M 1.00e-07	D-R -iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd
1								
22	marker_7.5	99	97	90	76	75	100	100
23	marker_7.0	97	91	75	50	83	87	86
24	marker_6.5	90	76	49	24	92	50	70
25	marker_6.0	75	50	23	9	92	23	57
26	marker_5.5	49	24	9	3	83	7	45
27	marker_5.0	23	9	3	1	75	0	38



FIG. 16D

A	B	C	D	E	F	G	H	I	J
Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10-6M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10-6M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10-7M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10-7M 1.00e-07	D-R -iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd	interp -log IC50 by Cmpd	est IC50 μM by Cmpd
1									
2	Cmpd20	101	108	100	75	100	100		<0.1
3	marker_7.5	99	90	76	75	100	100		
4	marker_7.0	97	75	50	83	87	86		
5	Cmpd07	100	48	43	83	77	79	6.78	0.17
6	Cmpd13	81	40	37	83	70	75	6.66	0.22
7	Cmpd15	99	46	36	83	70	75	6.66	0.22
8	Cmpd16	101	38	18	92	50	70	6.50	0.32
9	marker_6.5	90	49	24	92	50	70		
10	Cmpd17	81	47	24	92	43	67	6.38	0.41
11	Cmpd03	83	28	15	92	23	57	6.00	1
12	Cmpd05	89	21	22	92	23	57	6.00	1
13	Cmpd06	71	13	3	92	23	57	6.00	1
14	Cmpd09	68	22	15	92	23	57	6.00	1
15	Cmpd11	50	63	25	67	47	57	6.00	1
16	Cmpd18	79	22	32	92	23	57	6.00	1
17	marker_6.0	75	23	9	92	23	57		
18	Cmpd19	71	23	12	83	30	56	5.96	1.1
19	Cmpd10	61	24	5	83	20	51	5.75	1.8
20	Cmpd01	90	16	19	83	10	46	5.54	2.9
21	Cmpd04	70	24	29	83	10	46	5.54	2.9
22	Cmpd02	41	22	5	83	7	45		>3
23	Cmpd08	65	28	38	58	33	45		>3
24	Cmpd12	47	24	3	83	7	45		>3
25	Cmpd14	39	4	12	83	7	45		>3
26	marker_5.5	49	9	3	83	7	45		
27	marker_5.0	23	3	1	75	0	38		

FIG. 16E



40/64

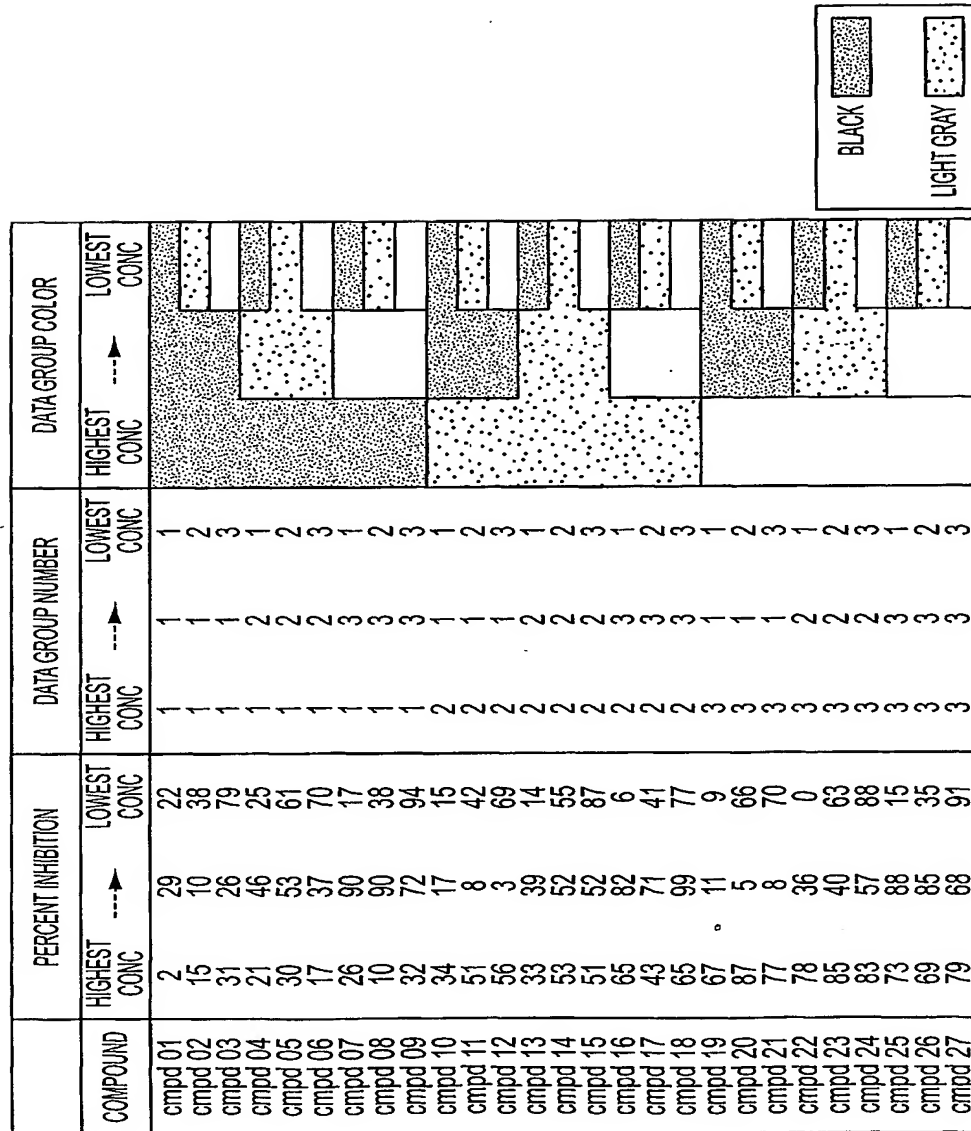


FIG. 16F

SUBSTITUTE SHEET (RULE 26)

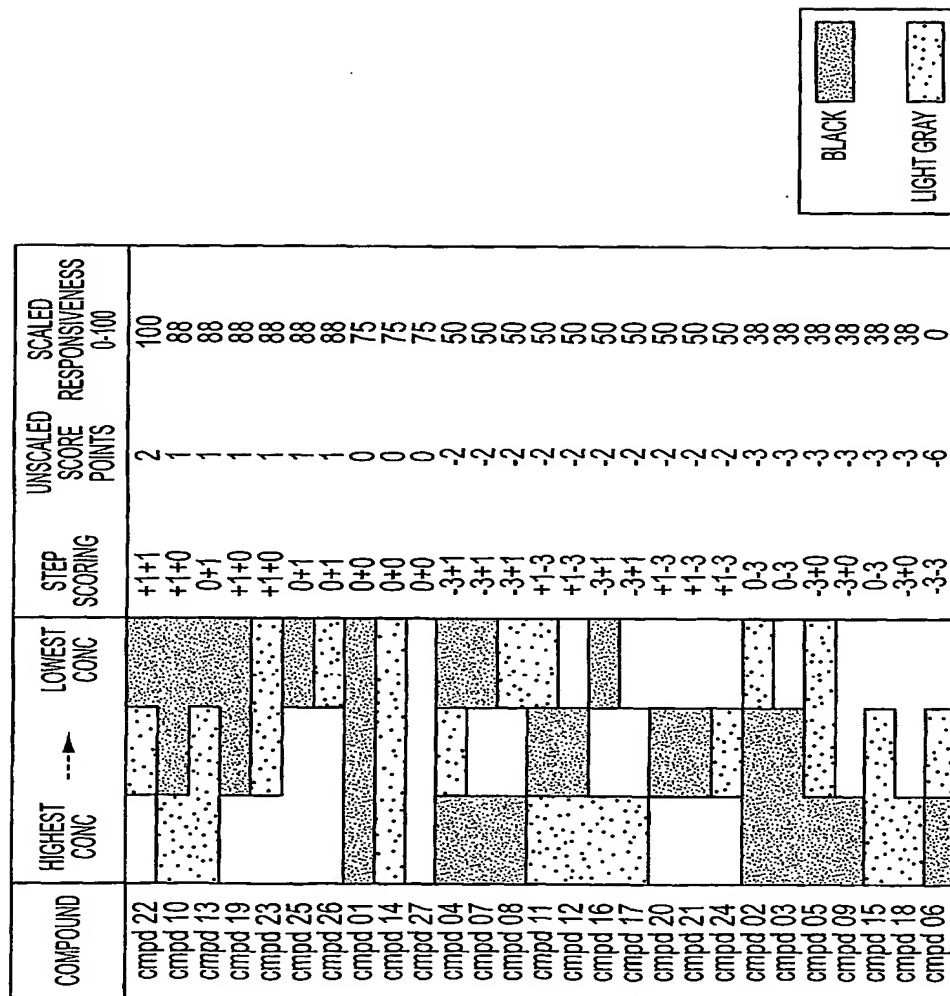
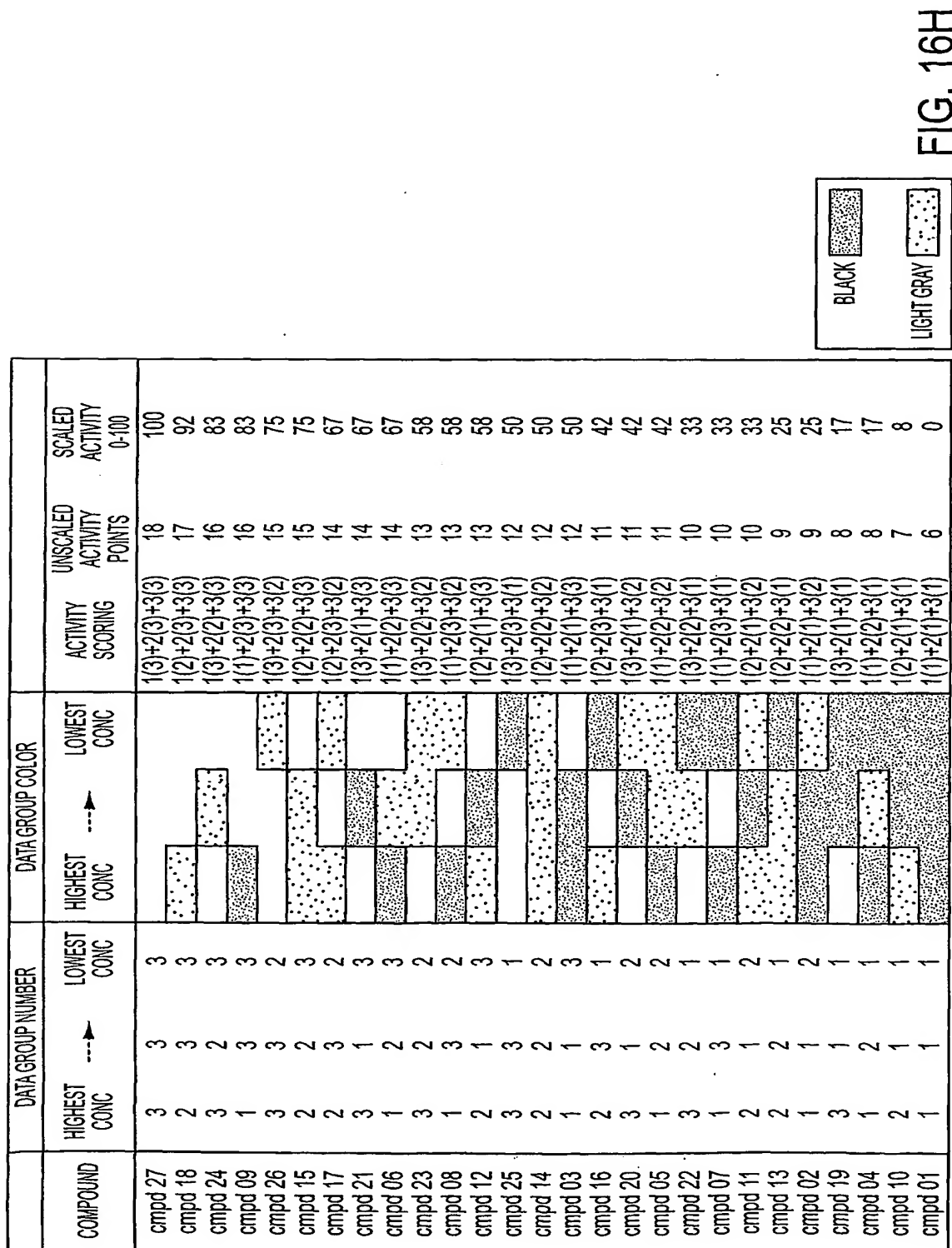


FIG. 16G

42/64



SUBSTITUTE SHEET (RULE 26)

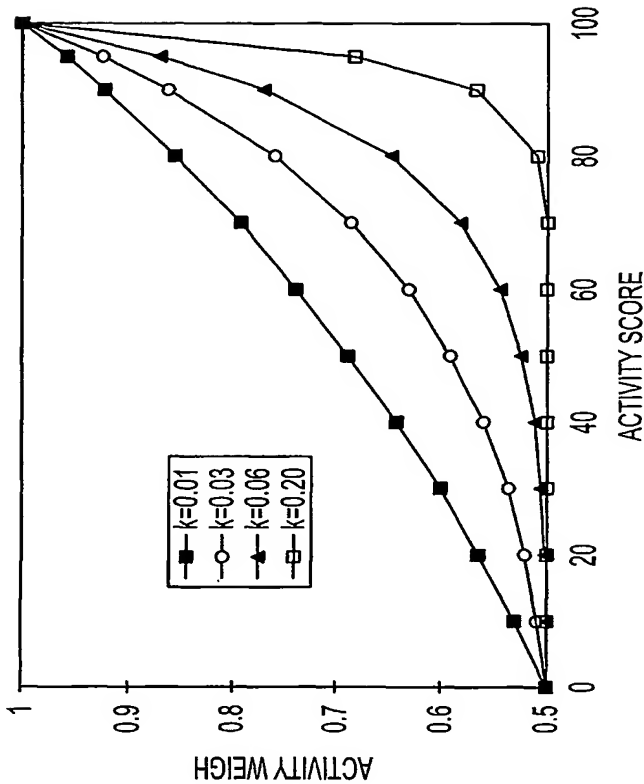


FIG. 16I

44/64

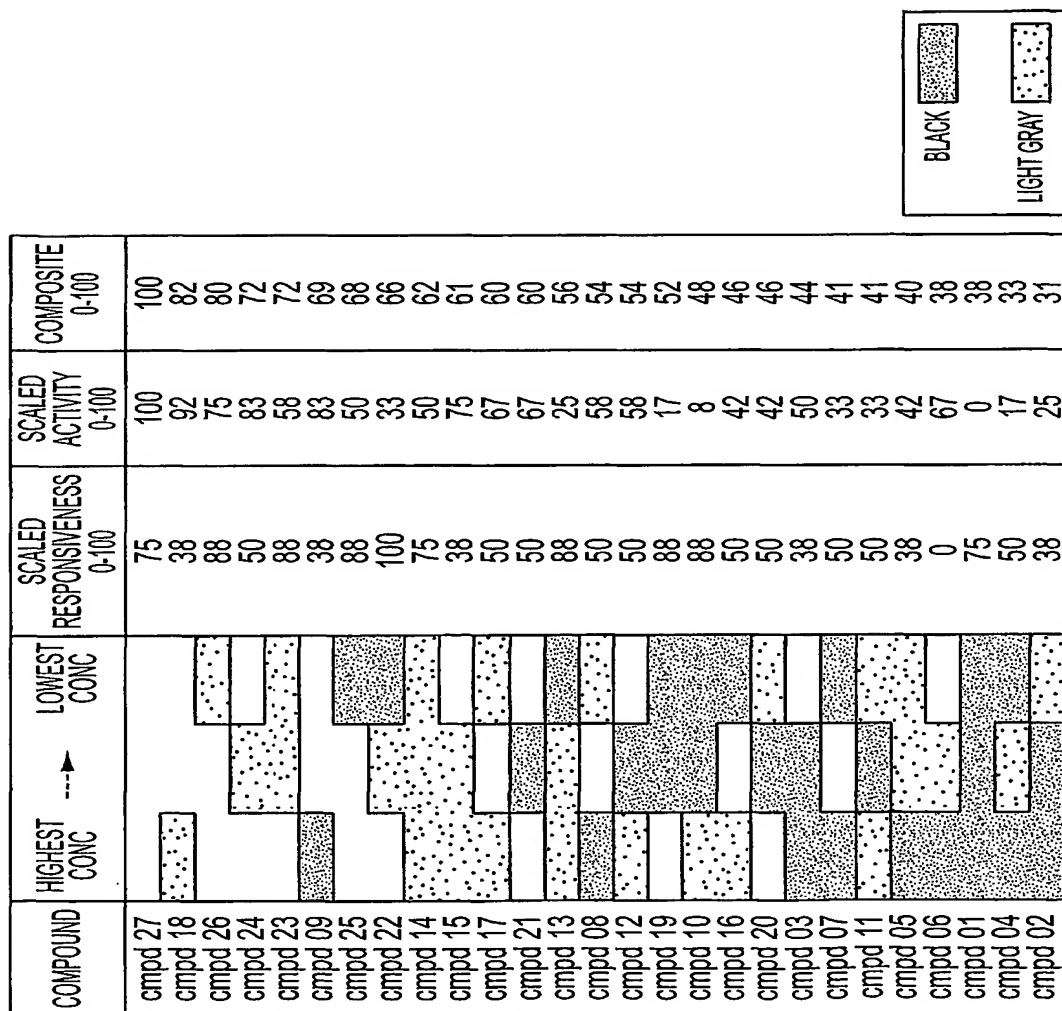


FIG. 16J

SUBSTITUTE SHEET (RULE 26)

45/64

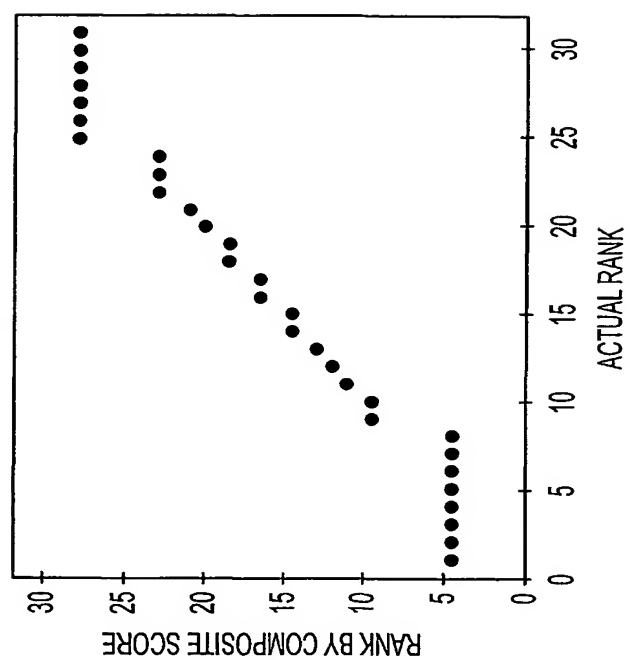


FIG. 16K

46/64

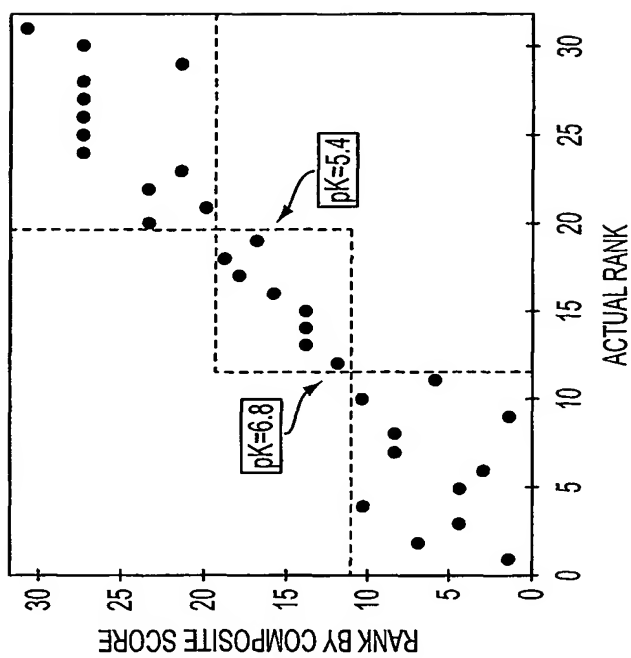


FIG. 16L

47/64

COMPOUND	% Inhib @ 3.00e-06	% Inhib @ 1.00e-06	% Inhib @ 3.00e-07	% Inhib @ 1.00e-07	D-R COMPOSITE SCORE	INTERP LOG IC50	est IC50 μM
M353875	100	102	83	75	79		<0.1
marker 7.5	99	97	90	76	79		
M221211	110	91	69	39	76	7.00	0.10
M371585	108	102	63	27	76	7.00	0.10
marker 7.0	97	91	75	50	76		
M345077	102	87	92	87	75	6.97	0.11
M371796	91	78	76	33	73	6.90	0.13
M143629	100	79	48	43	69	6.77	0.17
M371890	101	72	42	29	69	6.77	0.17
M371891	92	69	55	49	69	6.77	0.17
M309032	105	62	62	21	67	6.70	0.20
M198289	101	82	38	18	66	6.67	0.22
M224602	97	79	43	23	66	6.67	0.22
M318671	93	63	42	14	66	6.67	0.22
M273373	95	93	52	25	65	6.63	0.23
M371336	78	70	44	27	62	6.53	0.29
M371825	61	64	60	26	62	6.53	0.29
M181250	99	46	46	36	61	6.50	0.32
M338331	87	73	29	15	61	6.50	0.32
marker 6.5	90	76	49	24	61		
M143630	65	28	28	38	60	6.44	0.36

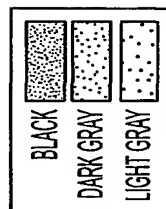


FIG. 16M

SUBSTITUTE SHEET (RULE 26)

48/64

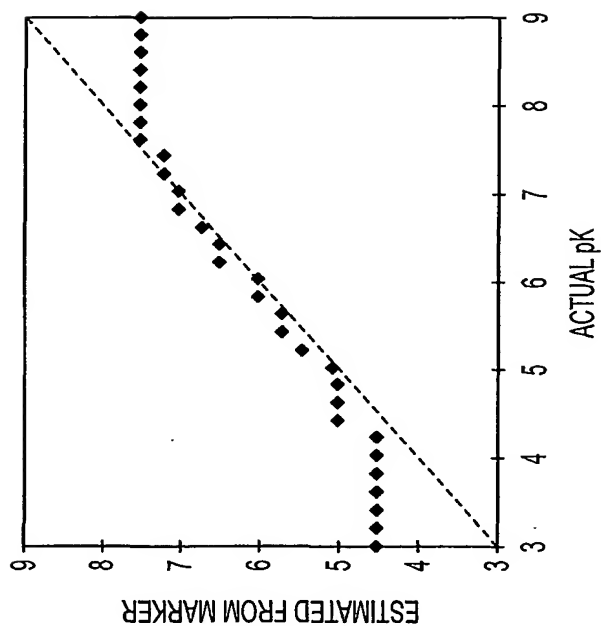


FIG. 16N

49/64

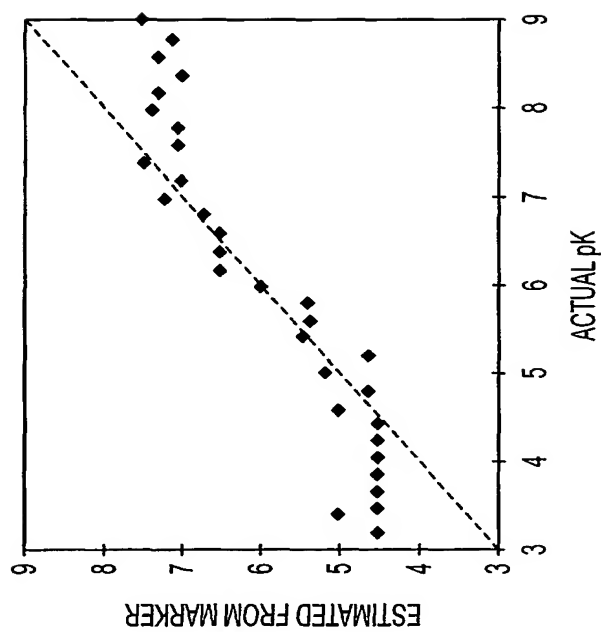


FIG. 16P

50/64

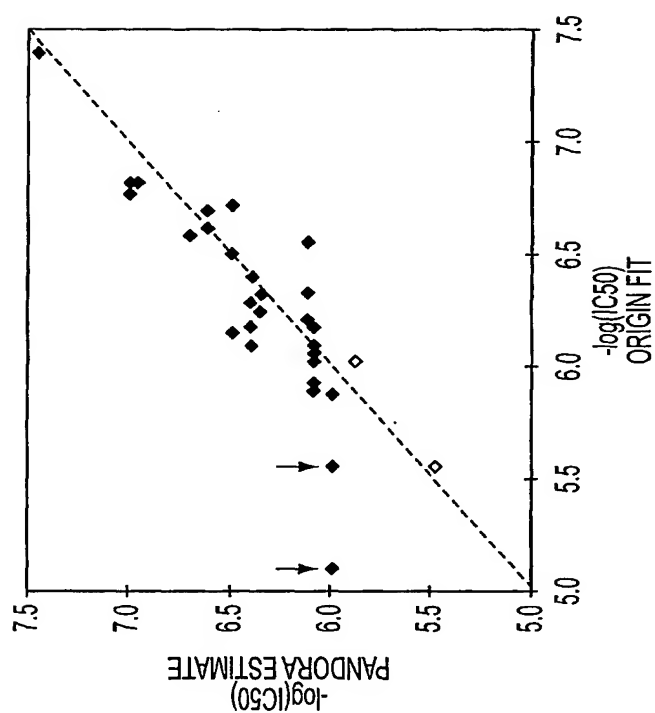


FIG. 16Q

	A	B	C	D	E	F	G	H	I
	Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 1x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M
1									
2	Cmpd01	N		29	30				
3	Cmpd02	N		42	5.5	41	3	22	5
4	Cmpd03	G		2.61	11	83	57	28	15
5	Cmpd04	N			30	70	25	24	29
6	Cmpd05	N		1.8	9.2	89	60	21	22
7	Cmpd06	D	8.86	6.5	3.7	71	41	13	3
8	Cmpd07	D	3.11	0.037	7.8	100	79	48	43
9	Cmpd08	D		0.089	N.A.	65	28	28	38
10	Cmpd09	D	0.119			68	41	22	15
11	Cmpd10	N	0.233			61	42	24	5
12	Cmpd11	N	4.31			50	77	63	25
13	Cmpd12	H	1.3	0.24		47	25	24	3
14	Cmpd13	H	1.17	0.194	30	81	59	40	37
15	Cmpd14	H	0.26	0.41		39	23	4	12
16	Cmpd15	H	0.369	0.148		99	46	46	36
17	Cmpd16			0.87	30	101	82	38	18
18	Cmpd17	K		0.223	N.A.	81	64	47	24
19	Cmpd18		5.27			79	54	22	32
20	Cmpd19		0.134			71	71	23	12
21	Cmpd20			0.317		101	109	108	100
22	Cmpd21	K		2.21		87	70	31	13
23	Cmpd22	B		0.15		94	77	36	12
24	Cmpd23	B				96	61	36	12
25	Cmpd24	B	3.487	0.27	0.4	110	91	69	39

FIG. 17A

52/64

* THIS COLORING INDICATES A DATA COLUMN WITH MIXED DATA TYPES.												
orig col	heading	# numeric	# text	# date	# blank	# total (longest col)	last occupied row num.	minimum (4 sig fig)	maximum (4 sig fig)	mean (4 sig fig)	standard dev (4 sig fig)	unique text strings and counts
A	Cmpd		24			24	25					(24 different)
B	Series		20		4	24	25					B(3) D(4) G(1) H(4) K(2) N(6)
C	Test1	12			12	24	25	0.119	8.86	2.385	2.726	
D	Test2	17			7	24	25	0.037	42	5.122	11.77	
E	Test3	10	2		12	24	25	0.4	30	15.76	12.59	N.A. (2)
F	HTS SPA Dose-Resp % Inhib @ 3x10-6M	23			1	24	25	39	110	77.57	20.43	
G	HTS SPA Dose-Resp % Inhib @ 1x10-6M	23			1	24	25	3	109	55.87	25.24	
H	HTS SPA Dose-Resp % Inhib @ 3x10-7M	23			1	24	25	4	108	35.52	21.85	
I	HTS SPA Dose-Resp % Inhib @ 1x10-7M	23			1	24	25	3	100	23.91	20.74	



FIG. 17B

A	B	C	D	E	F	G
project name	most important factor scored by Mngr A	most important factor scored by Mngr B	most important factor scored by Mngr C	less important factor scored by Mngr A	less important factor scored by Mngr B	less important factor scored by Mngr C
1						
2	2	3	2	1	2	2
3	1	1	1	2	1	2
4	1	1	3	1	2	3
5	3	3	3	1	2	1
6	3	3	3	3	1	2
7	2	3	1	3	2	1
8	3	1	1	3	3	1
9	3	1	2	3	1	3
10	3	3	1	3	2	1
11	2	1	3	1	1	1
12	1	1	1	2	1	1
13	3	3	1	1	3	2
14	3	1	3	3	3	2
15	3	2	2	3	3	3
16	2	3	2	2	2	2
17	1	2	2	1	2	3
18	2	1	3	2	2	2
19	3	3	3	1	3	3
20	2	2	2	3	1	1
21	3	2	3	1	3	2

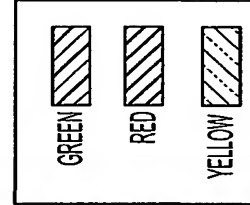
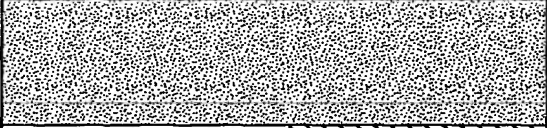
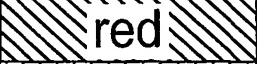

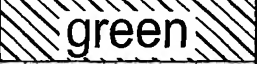
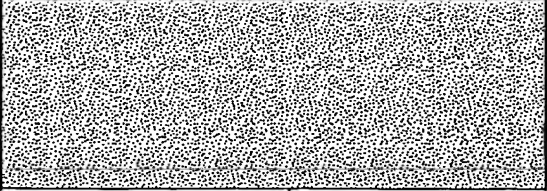


FIG. 18A

Click here to run these	
sheet	Portfolio
column(s)	B:G
# of colors	3
break 1	1
break 2	2
break 3	3
	
color 1	 red
color 2	 yellow
color 3	 green
	
Re-scale all?	





BLACK	
GREEN	
RED	
YELLOW	

FIG. 18B

SUBSTITUTE SHEET (RULE 26)

[illegible]





BLACK	
GREEN	
RED	
YELLOW	

FIG. 18C

	A	B	C	D	E	F	G	H
	project name	most important factor scored by Mngr A	most important factor scored by Mngr B	most important factor scored by Mngr C	less important factor scored by Mngr A	less important factor scored by Mngr B	less important factor scored by Mngr C	score (0-100)
1								
2	Proj 18	3	3	3	1	3	3	94
3	Proj 05	3	3	3	3	1	2	92
4	Proj 04	3	3	3	1	2	1	86
5	Proj 14	3	2	2	3	3	3	83
6	Proj 20	3	2	3	1	3	2	83
7	Proj 13	3	1	3	3	3	2	81
8	Proj 09	3	3	1	3	2	1	75
9	Proj 12	3	3	1	1	3	2	75
10	Proj 15	2	3	2	2	2	2	75
11	Proj 01	2	3	2	1	2	2	72
12	Proj 08	3	1	2	3	1	3	69
13	Proj 06	2	3	1	3	2	1	67
14	Proj 17	2	1	3	2	2	2	67
15	Proj 19	2	2	2	3	1	1	64
16	Proj 07	3	1	1	3	3	1	61
17	Proj 03	1	1	3	1	2	3	58
18	Proj 10	2	1	3	1	1	1	58
19	Proj 16	1	2	2	1	2	3	58
20	Proj 02	1	1	1	2	1	2	39
21	Proj 11	1	1	1	2	1	1	36

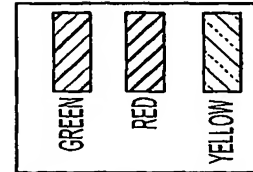
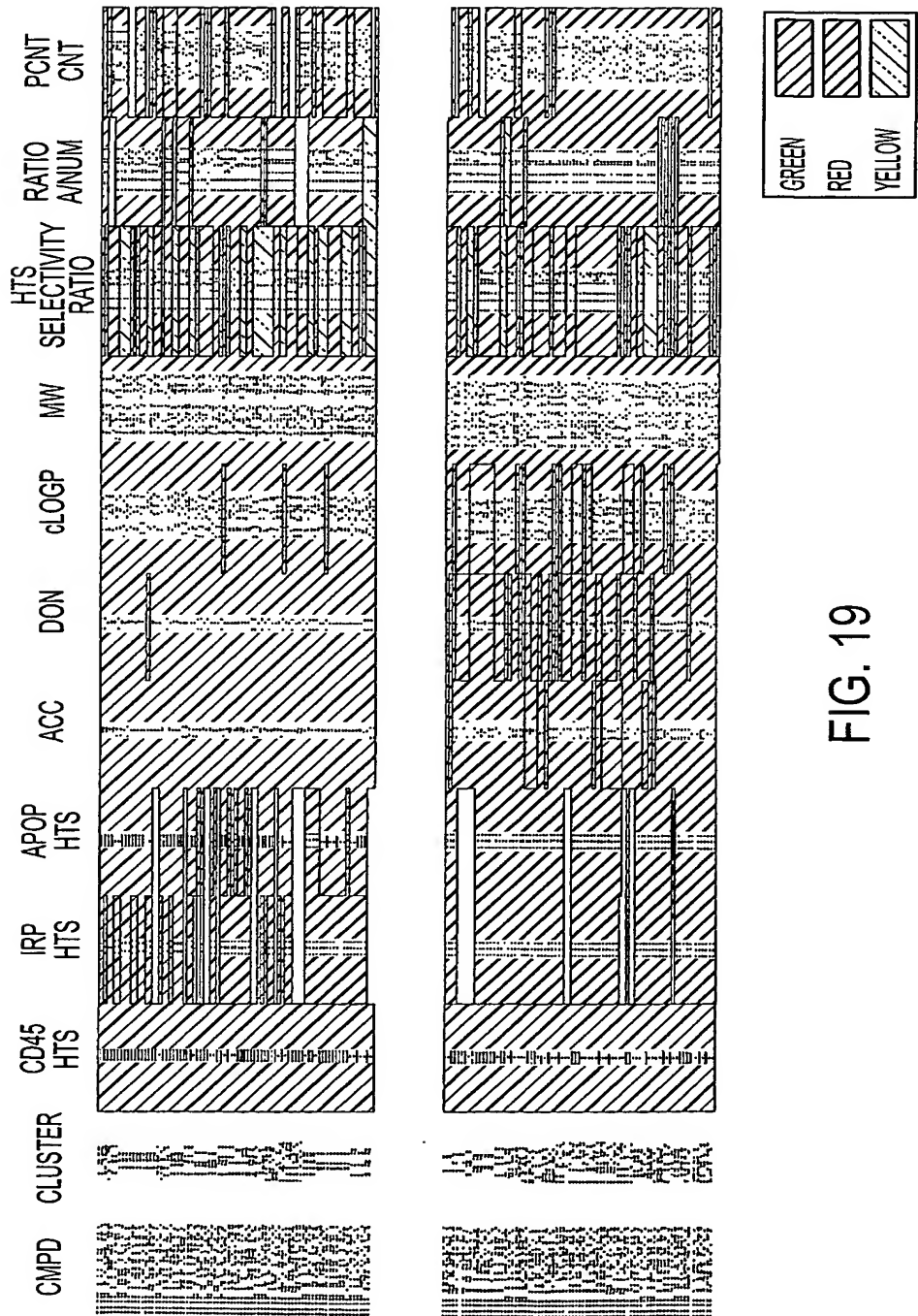


FIG. 18D

SUBSTITUTE SHEET (RULE 26)



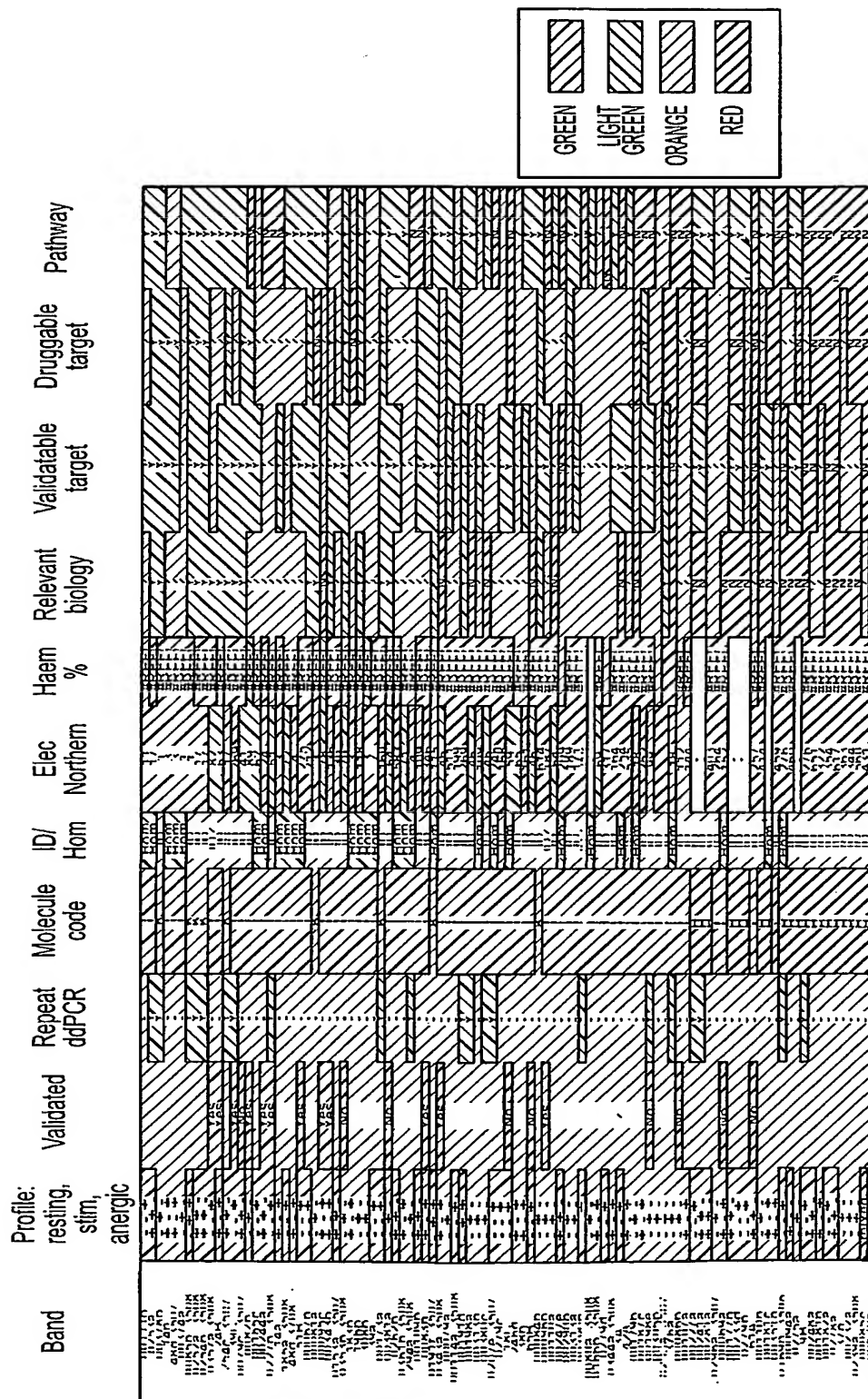


FIG. 20

project name	most important factor scored by manager 1	most important factor scored by manager 2	most important factor scored by manager 3	less important factor scored by manager 1	less important factor scored by manager 2	less important factor scored by manager 3
Proj 18						
Proj 05						
Proj 04						
Proj 14						
Proj 20						
Proj 13						
Proj 09						
Proj 12						
Proj 15						
Proj 01						
Proj 08						
Proj 06						
Proj 17						
Proj 19						
Proj 07						
Proj 03						
Proj 10						
Proj 16						
Proj 02						
Proj 11						

GREEN

RED

YELLOW

FIG. 21

Company	Disease 1	Disease 2	Disease 3	Disease 4	Disease 5	Disease 6	Disease 7	Disease 8	Disease 9	Disease 10	Disease 11	Disease 12
Company 1	Phase I		Phase II		LO/DE				LO/DE	Phase I		Phase I
Company 2	Phase I	LI	Phase II			LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	Phase II	
Company 3	LI/LO				LI/LO	Rx	LI/LO		Rx	Rx		
Company 4	LO				LO/Ph	Phase II	LO	Phase II			LO	
Company 5	LI/LO			LO	LO	Phase I	LO/DE			LO		HTS/II
Company 6	LO		Rx	Rx				Rx				
Company 7	LI/LO							Phase I	LO/DE	LI/LO	Phase I	LI
Company 8	LO/DE	Phase II			LO				LO		LO	
Company 9	LI/LO					Phase I	LI/LO	LI/LO	LI/LO		LI/LO	LI/LO
Company 10		LO		LO	Phase II					Phase II		
Company 11	LO			LO	LO							
Company 12	LO		Phase II		LO							LI/LO
Company 13	LI				Phase III		Phase III					LI
Company 14	LI/LO		LI/LO		LI/LO			LI/LO			LI/LO	
Company 15	LI		LI	LI	LI					LI		
Company 16	LI		LI	LI	LI		Phase II					
Company 17	Phase II				LI/LO	LI/LO						
Company 18			Phase II					Phase I				
Company 19	LI		LI				LI					
Company 20				LI/LO								
Company 21	LI/LO				Phase III							
Company 22	HTS		HTS			HTS				HTS		
Company 23	LI				LO/DE							
Company 24	LI											
Company 25	LI/LO											

CYAN	MAROON	ORANGE	RED	YELLOW
------	--------	--------	-----	--------

FIG. 22

SUBSTITUTE SHEET (RULE 26)

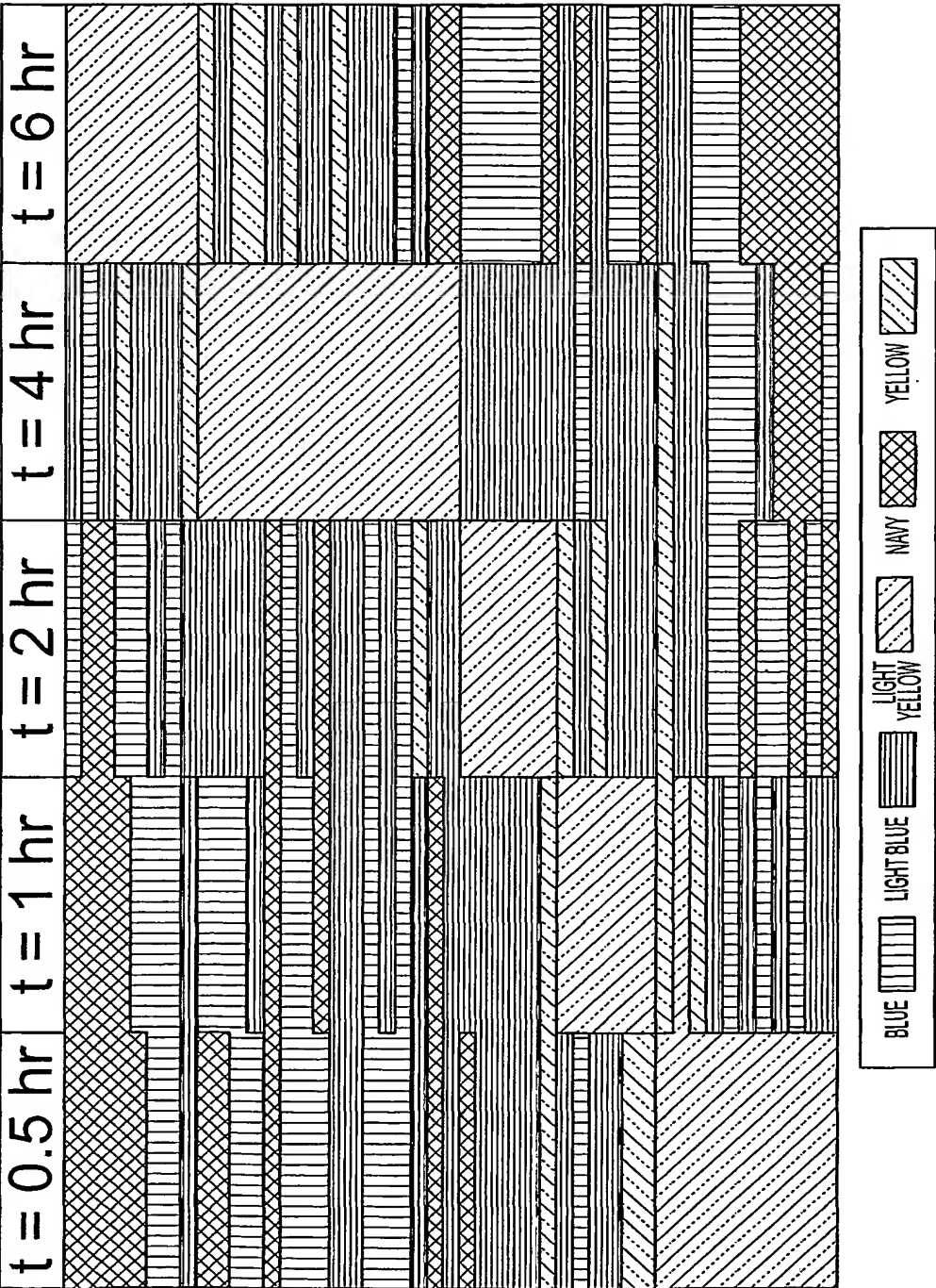


FIG. 23

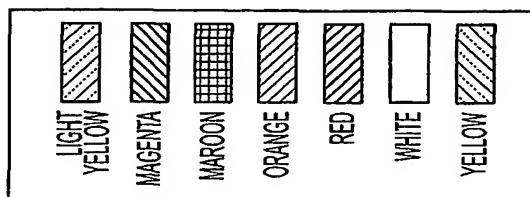
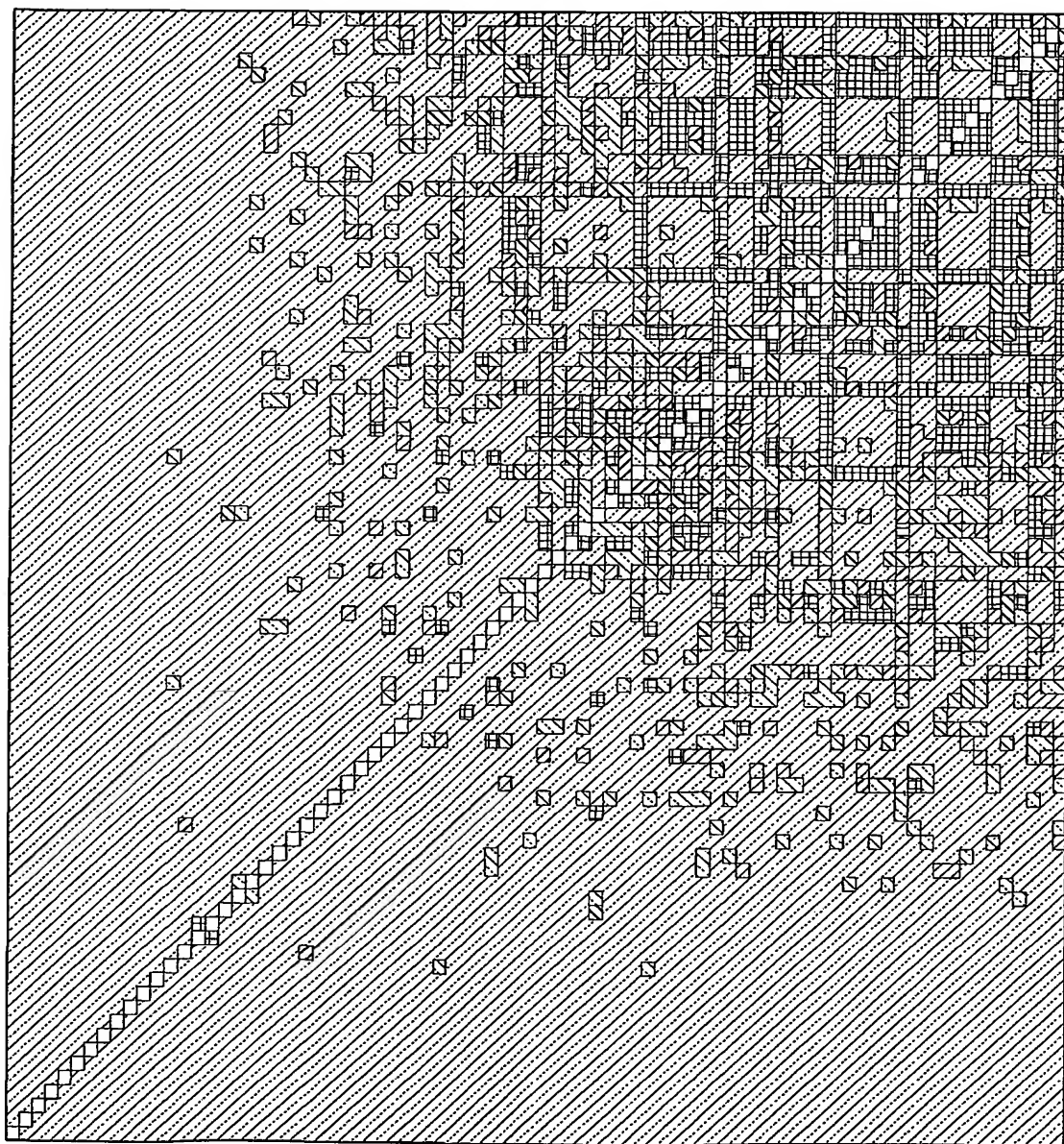


FIG. 24



SUBSTITUTE SHEET (RULE 26)

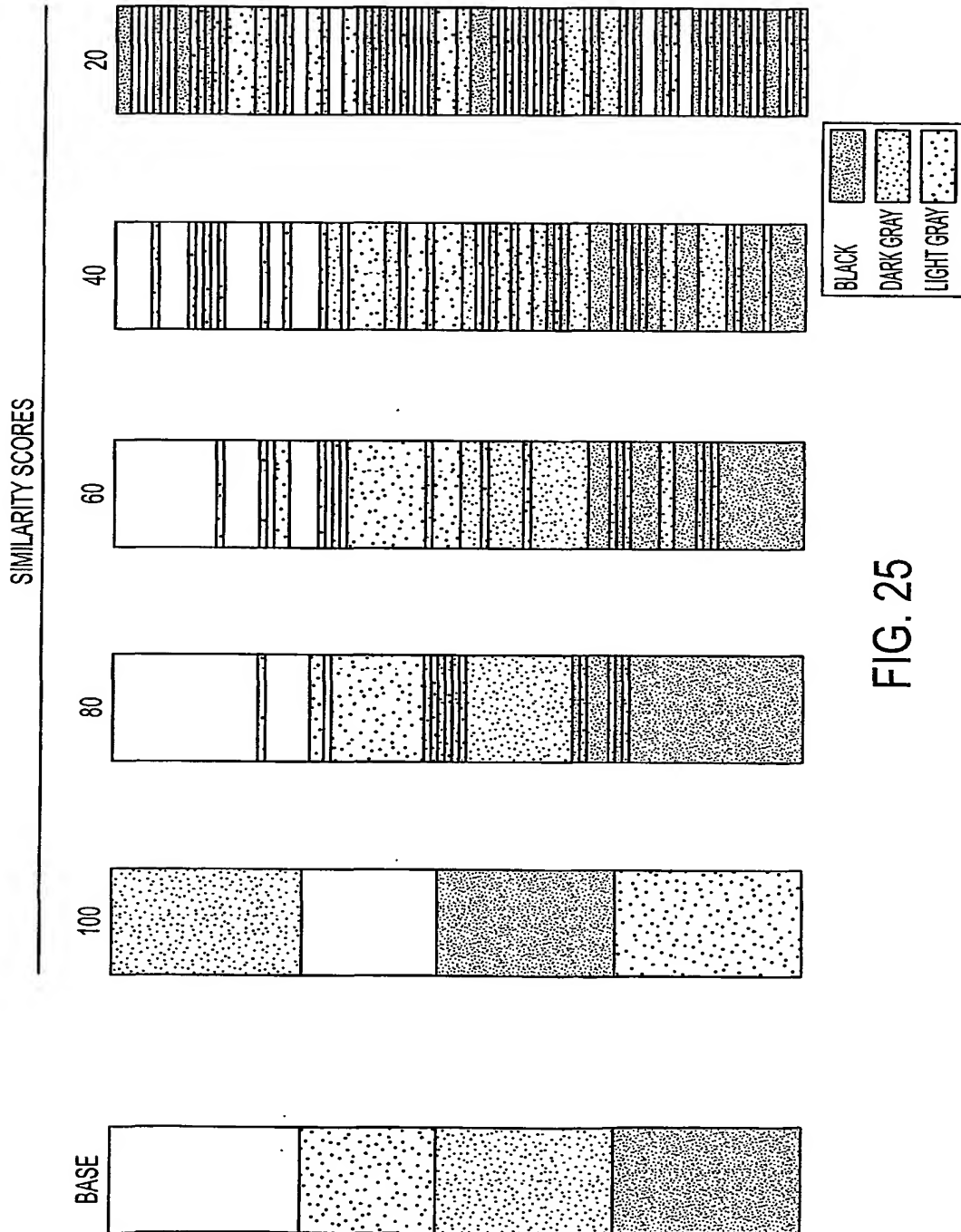


FIG. 25

SUBSTITUTE SHEET (RULE 26)

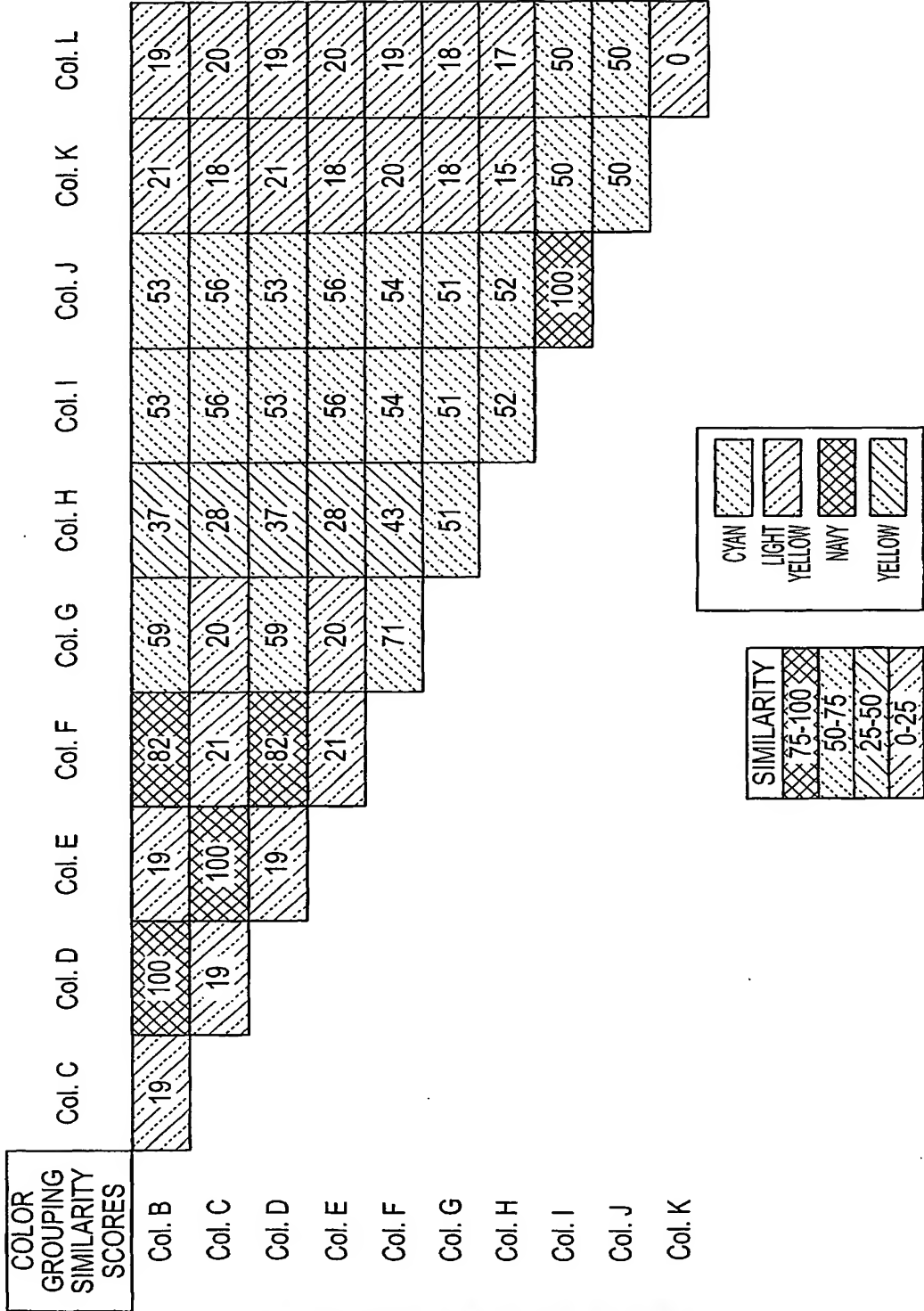


FIG. 26

SUBSTITUTE SHEET (RULE 26)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 00/20401

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06K9/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, COMPENDEX, INSPEC, WPI Data, IBM-TDB, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>ANONYMOUS: "Dynamic Layout Mechanism for the Massive-Node Server Status Monitor"</p> <p>IBM TECHNICAL DISCLOSURE BULLETIN, vol. 36, no. 5, 1 May 1993 (1993-05-01), pages 169-170, XP000408951</p> <p>New York, US</p> <p>the whole document</p> <p style="text-align: center;">---</p> <p style="text-align: center;">-/--</p>	1-48

☒ Further documents are listed in the continuation of box C.

☐ Patent family members are listed in annex.

* Special categories of cited documents:

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

- *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- *G* document member of the same patent family

Date of the actual completion of the international search

12 January 2001

Date of mailing of the international search report

22/01/2001

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Granger, B

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 00/20401

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EISEN M B ET AL: "Cluster analysis and display of genome-wide expression patterns" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA,US,NATIONAL ACADEMY OF SCIENCE. WASHINGTON, vol. 95, December 1998 (1998-12), pages 14863-14868, XP002140966 ISSN: 0027-8424 page 14863, right-hand column, paragraph 3; figure 2	1-48
A	STANTON D T ET AL: "Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery" JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, JAN.-FEB. 1999, ACS, USA, vol. 39, no. 1, pages 21-27, XP000971515 ISSN: 0095-2338 the whole document	1-48

This Page Blank (uspto)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☒ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☒ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☒ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

This Page Blank (uspto)